Cluster Randomized Trials

10

Michael J. Campbell

Contents

M.J. Campbell
Medical Statistics Group, Health Services Research, ScHARR, University of Sheffield, Sheffield, UK

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_47,

10.1 Introduction

10.1.1 Basics

A cluster randomized trial is one in which groups of subjects are randomized rather than individuals. They are sometimes known as group randomized trials. This chapter will describe the design and analysis of such trials. Examples of cluster trials in health are given in Box 10.1.

Cluster trials are used widely in the evaluation of interventions in health services research. They can be divided into two types. The first type is exemplified by the first three rows in Box 10.1: community randomized trials where the clusters are complete communities (some authors call these "large field trials"). These are generally characterized by a relatively small number of clusters each enrolling a large number of subjects. The aim of the trial by Grosskurth et al. (1995) cited in Hayes and Moulton (2009) was to reduce the prevalence of HIV infection by treating other sexually transmitted diseases. It involved six intervention communities and six matched control communities. In each, a random sample of 1,000 adults was selected in each community and followed up for 2 years to measure the incidence of HIV infection. The trial COMMIT (Gail et al. 1992) was to test an intervention aimed at communities to encourage citizens to stop smoking. It had 11 matched pair clusters. The sex-education trial by Wight et al. (2002) randomized 25 schools, 13 into intervention and 12 control, and interviewed all 13- to 14-year-olds at the schools and the same children after 2 years.

The second type of cluster trial is closer in design to an individually randomized trial. It typically uses more clusters and relatively smaller cluster sizes. Examples of "small cluster size" trials are given in the second half of Box 10.1. As an example of the second type, consider in more detail the DESMOND trial described by Davies et al. (2008) (DESMOND – Diabetes Education and Self Management Ongoing and Newly Diagnosed), which involved 105 general practices in the intervention and 102 in the control. The purpose of the trial was to investigate whether an intensive education package can be used to reduce glycosylated haemoglobin (HbA1c%) in patients who have type II diabetes. In the UK diabetes is usually treated in primary

Box 10.1. Examples of cluster trials

Unit	Intervention	Example
Rural communities	Treatment of coexisting disease	Grosskurth et al. (1995)
Communities	Education	Gail et al. (1992)
Schools	Education packages	Wight et al. (2002)
Groups	Diabetes education	Davies et al. (2008)
Doctors	Patient-centered care	Kinmonth et al. (1998)
Patients	Teeth fillings	Soncini et al. (2007)

care, and it was deemed impossible to randomize people in the same practice to different treatments. Thus practices were chosen (at random) as either "intervention" practices or "control" practices. DESMOND is usually taught as a course to groups of eight people at the same time, so the course was the cluster in this case. Kinmonth et al. (1998) randomized general practitioners into those who would receive training in "patient-centered 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-centered care" to "paternalistic" care with successive patients. The outcome was measured by HbA1c% in their diabetic patients. Soncini et al. (2007) looked at the survival of amalgam versus composite fillings in teeth and randomized 267 children into each group. It was deemed simpler to ensure each child either had amalgam or composite fillings and so survival times of the fillings will be clustered by mouth.

The main reason for using a cluster trial is fear of contamination. This occurs when subjects in the control group are exposed to the intervention. Thus people living in the same community could not fail to notice a mass education program delivered on the television or local newspaper. In the DESMOND trial, patients may wonder why people with the same doctor were getting different treatment and demand the same for themselves. Doctors trained in a new technique will find it difficult to revert to an old technique at the 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. Another reason is that it may be more effective or cheaper to deliver an intervention to a group. For example, patients in the same education program 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 the trainer. A third reason for adopting a cluster design is administrative convenience or necessity; it is often easier to deliver an intervention to a group of people when it may involve an expensive piece of equipment or training health professionals or it may be impossible to randomize individuals. Again, this was true of DESMOND, where it was much cheaper to deliver the intervention in groups. Sometimes it appears easier to get ethical consent when all of a group are getting the intervention.

The most important point, with regard to the analysis, is that observations are not independent. Observations within a particular cluster are correlated, and although this correlation may be weak, it can have a major effect on the analysis as we shall see.

It is worth defining a few terms. The *intraclass correlation (ICC)* is the ratio of the between cluster variance to the total variance of an outcome variable and is often denoted by ρ. Different designs can lead to different formulas for estimating ρ. A simple method is described in the next section. 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

$DE = 1 + (m - 1)\rho$. This is also called the *sample size inflation factor (SSIF)* in chapter ▸Generalized Estimating Equations of this handbook, since the same factor that inflates the variance will also inflate the required sample size. Extensions of this formula to the case of variable cluster sizes are given in Sect. 10.2.2.3.

10.1.2 Looking at Data

Figure 10.1 shows the HbA1c% in diabetic patients after 1 year from randomization by practice and by intervention/control from the DESMOND trial (Davies et al. 2008).

We are interested in the difference in the mean HbA1c% for intervention and control. However, one can see that there is a good deal of variation within practices but that some practices have in general high values and some practices have 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. The intraclass correlation ρ is a measure of how much subjects within a cluster are correlated. It is the ratio of the between cluster component of variance σ_B^2 to the total variance $\sigma_B^2 + \sigma_W^2$ where σ_W^2 is the variance within clusters. We can estimate these using a simple analysis of variance. This is shown in Table 10.1 as the

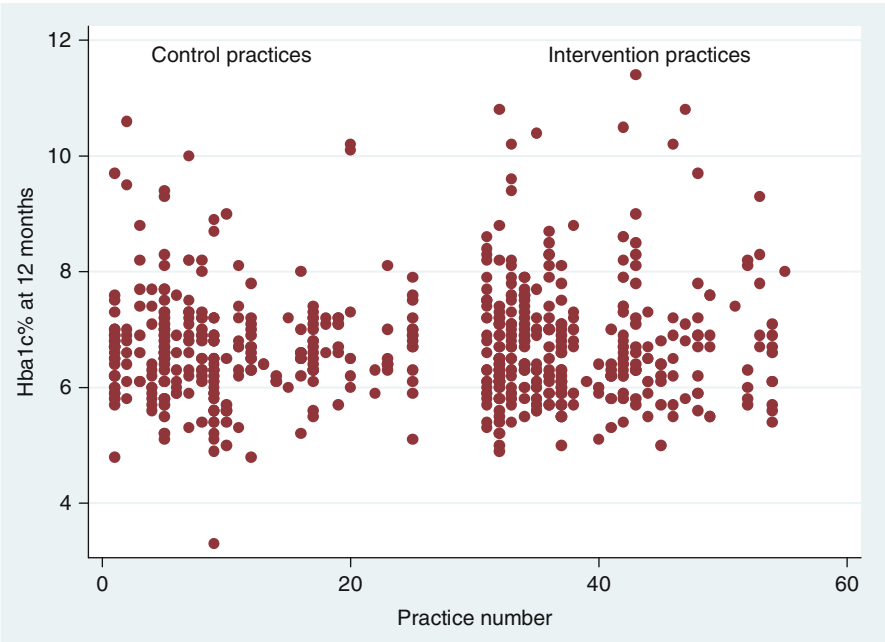

Fig. 10.1 HbA1c (%) at 1 year by control/intervention from DESMOND (Davies et al. 2008)

Table 10.1 Output showing ANOVA to estimate the ICC for data from DESMOND using Stata v11

```
Loneway hba1c12 practice
```

Source	SS	df	MS	F	Prob > F
Between practice	77.357747	46	1.6816902	1.84	0.0009
Within practice	539.41716	589	.91581861		
Total	616.77491	635	.97129907		

Intra-class correlation	Asy. S.E.	[95% Conf.	Interval]
0.05914	0.02832	0.00364	0.11464

output from the Stata command "loneway" (Statacorp 2009). We have 636 subjects in 47 practices. Here, σ_W^2, the *within practice mean square* (*MS*), is 0.9158. For a fixed cluster size, m, we can estimate the between cluster component of variance from the fact that the *between practice* $MS = \sigma_B^2 + m\sigma_W^2 = 1.6817$. Since here m is about 13.5, we find that $\sigma_B^2 = 0.0567$ and so ρ is $0.0567/(0.0567 + 0.9158) = 0.0583$. The program gives $\rho = 0.0591$, which is slightly different since it takes into account variable cluster size. It is important to appreciate that this procedure gives reasonable values for ρ even for binary outcomes, since it is a moment estimator and does not require distributional assumptions.

10.1.3 Overview of Chapter

Section 10.2 is concerned with the design of cluster randomized trials and how to estimate the number of patients and the number of clusters required. Section 10.3 discusses the analysis and presentation of such trials. Section 10.4 discusses other considerations for cluster trials and software for their analysis. Section 10.5 concludes the chapter and suggests further reading.

10.2 Design of Cluster Randomized Trials

10.2.1 Cohort Versus Cross-Sectional Designs

Many community intervention trials are longitudinal in nature, allowing a choice between a cohort design and a cross-sectional design. For a cohort design, clusters are randomly assigned to intervention groups, with or without stratification. Cohorts sampled from each cluster are then measured over two or more time points, with at least the first measurement occurring before randomization. They are useful

in looking at how the intervention changes the health or behavior of individual subjects. The baseline and follow-up subjects are the same people. People who drop out will often be different from those that stay and so the follow-up group may not be typical of the whole population from which the cohort was chosen. Thus it is important to report drop-out rates and do sensitivity analyses to consider whether the nature of the drop-outs may affect the conclusions.

In contrast a cross-sectional design involves randomizing large groups of subjects, such as towns. A random sample is taken before and a random sample taken after the intervention with similar samples taken in the control population. Thus the subjects before and after the intervention are not necessarily the same. On the one hand, a cross-sectional study should be a representative sample of the population, since it is based on a random sample. On the other hand, 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. Cross-sectional designs are useful when the main focus is on change in behavior or health in a community. It can be helpful in some situations to ask those after the intervention whether they were aware of it. For example, in an evaluation of a government advertising campaign, we asked subjects after the campaign if in fact they had seen it (Mills et al. 1986). Only 31% of the sample were, in fact, aware of the campaign, which may partly explain its lack of effectiveness.

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 one. However, Feldman and McKinlay (1994) 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 aging of the cohort over time. Assuming that changes related to the aging 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 by Nixon and Thompson (2003), who described and compared several approaches that might be taken to the analysis of repeated cross-sectional samples.

10.2.2 Power and Sample Size

10.2.2.1 Number of Clusters and Number of Subjects per Cluster

In cluster randomized trials, there are two sample size choices to be made: the number of clusters and the number of subjects per cluster. The usual situation is where the cluster size is fixed, and to increase the power we need to increase the number of clusters.

Suppose we needed n' subjects in an individually randomized trial to detect an effect size δ with two-sided significance α and power $1-\beta$ (e.g., using the tables in Machin et al. 2008). Then to allow for clustering where we have equal clusters of size m, we should increase the sample size (using *VIF/SSIF*) to n where n is given by

$$n = n'(1 + (m-1)\rho) \tag{10.1}$$

to achieve the same power and significance level (Hsieh 1988). The number of clusters is determined by $k = n/m$.

An alternative situation is where the number of clusters is fixed, and one wishes to determine the number of subjects per cluster. Since the design effect requires knowledge of the number of subjects per cluster, m, one has to guess m first, to find n and then recalculate m from $m = n/k$ and then reiterate. A simpler solution is to use the fact that (Campbell 2000)

$$m = \frac{m'(1-\rho)}{1-m'\rho}, \tag{10.2}$$

where $m' = n'/k$ is the number of subjects per cluster required before adjusting for clustering. Suppose that for a given effect size, significance level, and power, we require m' subjects per cluster in an individually randomized trial. If ρ is greater than $1/m'$, then m becomes negative, which is impossible and so one can never achieve the required power simply by increasing the number of subjects per cluster, and one will have to increase the number of clusters. Even if ρ is only slightly less than $1/m'$, the numbers per cluster become very large. Thus a useful rule of thumb for continuous outcomes is that the power does not increase appreciably once the number of subjects per cluster exceeds $1/\rho$. For example, if it is believed that the *ICC* is about 0.05, then it is not worth enrolling more than about 20 subjects per cluster for a continuous outcome. However, if the *ICC* of a continuous outcome is near 0.001, which is often typical of community intervention trials, then a sample of 1,000 subjects per cluster may be worthwhile, particularly if recruiting new clusters is difficult. Of course, with binary outcomes, when the incidence is low, large numbers of patients per cluster are required unless the effect of the intervention is very large.

Flynn et al. (2002) addressed the issue of whether it is worth recruiting an extra cluster, or to recruit more individuals to existing clusters. They showed how the use of contour graphs of power by number of clusters per treatment arm and cluster size can be usefully exploited. For example, consider a hypothetical trial in which

18 clusters have already been recruited in each of two treatment arms and in which at least 30 individuals can be recruited from each cluster. Suppose the *ICC* is 0.05 and the target standardized difference is 0.25. We then currently expect about 75% power. To achieve at least 80% power, we can show there are two options: (1) recruiting 20 extra individuals in each existing cluster; (2) recruiting two extra clusters in each arm. The question may then be which of these options is the least costly, and this would be the option to choose.

Although values of ρ in cluster randomized trials tend to be small (typically around 0.05 for primary care trials (Campbell 2000)) and in community randomized trials even smaller (usually less than 0.01 and often near 0.001, Donner and Klar 2000), the resulting inflation of the sample size may be very substantial when combined with clusters of large size. For example, in the trial of HIV reduction (Grosskurth et al. 1995), communities were on average 1,000 adults and so even an *ICC* as low as 0.0001 would have the effect of doubling the required sample size. Results from earlier studies in a specific setting of design effects likely to arise in cluster randomized trials implemented are also helpful to investigators. Gulliford et al. (2005) gave examples of variance components for some common outcomes which can be helpful for future planning.

The sample size formulas we give here assume that data for sample size estimation are obtained from a single sample of clusters from the population of interest, that is, that the intervention itself is not associated with the cluster size. The problem here is that variable cluster sizes will affect the power, and this will be discussed in Sect. 10.2.2.3. Examples where this is not true have been given by Campbell (2000). A particular example is Kinmonth et al. (1998) where the intervention was to train doctors in treating newly diagnosed diabetics, and these same doctors were the ones who diagnosed the diabetes and recruited the patients. They found the intervention clusters were larger than the control because the doctors who had received the intervention seemed more likely to find people with diabetes. This is known as recruitment bias and is a particular problem with cluster trial.

In summary, ignoring clustering effects in the design stage of a trial can lead to an elevated type 2 error, while ignoring it at the analysis stage inevitably leads to an elevated type 1 error. In other words, if an investigator ignores clustering when planning a study, the study is likely to be too small and so underpowered. If the investigator ignores clustering in the analysis, then the standard error of the estimate is likely to be underestimated, and so the observed p-value will be too low, and results declared significant at a given level, when in fact the null hypothesis should not have been rejected at that level.

10.2.2.2 Allowing for Imprecision in the *ICC*

Much of the sample size literature deals with the difficulty of obtaining accurate estimates of between community variation, and hence of ρ, that are needed for sample size planning.

In practice estimates of ρ for a given outcome variable are usually derived from previously reported studies using similar randomization units. However these estimates are frequently based on studies which themselves are of small size, and thus their inherent inaccuracy may lend the investigators a false sense of confidence.

Turner et al. (2004) have shown how to incorporate uncertainty in the *ICC* in a Bayesian framework to obtain an "average" power (cf. chapter ▸Bayesian Methods in Epidemiology of this handbook). They discussed the use of prior distributions for the ICC and showed how the uncertainty about this parameter can be expressed in the form of a parametric distribution which naturally leads to a distribution of projected power for any particular design. This Bayesian approach toward the determination of sample size could then be followed by a statistical analysis within the traditional frequentist framework. If the total sample size were fixed, and $n = m \times k$, it is better to increase the number of clusters k and have smaller cluster sizes. Increasing the number of clusters also considerably reduces the lower limit of the posterior distribution of power. In other words, uncertainty in the *ICC* will produce uncertainty in the actual power of a study, but a design based on a greater number of clusters has less chance of having a very low power. In general, one should try and recruit as many clusters as possible.

An alternative approach to dealing with uncertainty in observed values of ρ is described by Feng and Grizzle (1992), who proposed the use of a method similar in principle to the bootstrap procedure. For a simple discussion of the bootstrap in this context, see Carpenter and Bithell (2002). Their approach requires the simulation of results of studies of the same size to that which yielded the observed estimate. One then can substitute the values of ρ obtained from each simulation into the appropriate sample size formula to generate a distribution of projected powers, followed by the selection of a point on this distribution, for example, the 90th percentile, that reflects the degree of conservativeness desired.

10.2.2.3 Allowing for Varying Cluster Sizes

Variation in cluster size is another source of imprecision, and Kerry and Bland (2001) following Donner et al. (1981) suggested using

$$DE = 1 + (m_a - 1)\rho, \text{ where } m_a = \sum m_i^2 / \sum m_i . \tag{10.3}$$

The problem of using this formula is that the individual cluster sizes must be known prior to conducting the trial.

Eldridge et al. (2006) modified (10.3) to give

$$DE = 1 + ((cv^2 \frac{(k-1)}{k} + 1)\bar{m} - 1)\rho, \tag{10.4}$$

where $\bar{m}$ is the mean cluster size, s_m is the standard deviation ($s.d.$) of the cluster size given by $s_m = \sqrt{\sum (m_i - \bar{m})^2/(k-1)}$, and the coefficient of variation $cv = s_m/\bar{m}$. They suggested that formula (10.4) is more practical than (10.3) since the value of cv may often be known in advance. Eldridge et al. (2006) also provided some examples of the coefficient of variation for sample sizes typically seen in cluster randomized trials. Data from similar trials would be the first source of values for cv. Alternatively one could ask what is likely to be the maximum and minimum size cluster and estimate s_m as the range is divided by 4 (although strictly speaking this would require the data to be normally distributed).

Eldridge et al. (2006) showed empirically that as the average cluster size increases, the coefficient of variation tends toward 0.65. Primary care trials in the UK tend to have values of cv between 0.42 and 0.75. In the trial of Davies et al. (2008), the mean cluster size was about 14, and the $s.d.$ of cluster size was 12, suggesting a cv of about 0.86 which is somewhat more variable than many. The range of cluster sizes was 1 to 48, so the rule of thumb of estimating s_m as $(48-1)/4 = 11.75$ is quite accurate. We give an example of a sample size calculation in the next section.

10.2.2.4 Example of Effect of Variable Cluster Size

Suppose an investigator stated that the expected number of patients per cluster was 10 and the estimated intra-cluster correlation coefficient was 0.05. A preliminary sample size calculation showed that the estimated sample size required for a given power, significance level, and effect size without taking account of clustering was 200 patients and so 20 clusters. Then the estimated sample size required taking account of clustering but ignoring variation in cluster size is

$$n = (1 + (\overline{m} - 1)\rho)x\ 200 = 1.45 \times 200 = 290 \text{ patients } = 29 \text{ clusters.}$$

We can argue that a conservative estimate of the expected minimum size of a cluster is 1 patient (no cluster can have less than 1) and the expected maximum is 30 (since we would stop recruiting above this level). The cv is then estimated $(29/4)/10 = 0.725$. From Eq. 10.4, the design effect

$$DE = 1+((1+cv^2\times 28/29)\overline{m}-1)\rho = 1+((1+0.725^2\times 0.966)10-1)0.05 = 1.70.$$

Thus we would need $1.70 \times 200 = 340$ patients $= 34$ clusters, an increase of 5 clusters to allow for cluster size variable.

10.2.2.5 Sample Size Re-Estimation

When we do not know the value of the variance components required to ascertain a sample size, one suggestion is to conduct a pilot study to provide these estimates (Friede and Kieser 2006). We can then obtain an estimate of how many more patients we will need to recruit and use the patients from the pilot in the final analysis (an internal pilot). This procedure has been extended to cluster randomized trials by Lake et al. (2002). With this approach, we conduct a pilot cluster trial, and at an interim point in the study, several nuisance parameters, including ρ, the mean cluster size and measures of cluster size, variation, are estimated, followed by re-estimation of the final required trial size. Although this procedure is most suited to trials that randomize a relatively large number of clusters, such as families or households, over an extended period of time, Lake et al. (2002) pointed out that it could also be applied to at least some community intervention trials provided the participating clusters are recruited prospectively. However, Turner et al. (2004) showed that imprecision in the estimate of ρ is not accounted for in this application and suggested their Bayesian method could easily be extended to do so. More experience on the application of internal pilot studies to such trials is clearly needed.

10.2.2.6 Adjusting for Covariates

Sometimes the reason for a positive *ICC* is that subjects within a cluster have similar covariates. Several authors have shown that adjusting for covariates either at the community or individual level can improve the power of a trial by reducing the magnitude of the between cluster variation (Campbell 2000; Feng et al. 1999). Additional gains in power may be realized by modeling individual level covariates (e.g., age) and also cluster level covariates (e.g., mean cluster age) as described by Klar and Darlington (2004). Moerbeek (2006) suggested that often a cheaper strategy than recruiting another cluster would be to measure additional covariates related to the outcome in order to reduce the variance. This, of course, requires the cost of measuring the covariate to be relatively small and the correlation of the covariate and the outcome to be reasonably high.

10.2.3 Matched Pair Trials

10.2.3.1 Design of Matched Pair Studies

Because cluster trials involve randomizing relatively low numbers of groups, we cannot rely on randomization to ensure balance between treatment arms in important prognostic variables. A common technique to try and ensure balance is to match clusters into pairs and then randomly allocate one member of each pair to the intervention and one to the control.

Matched pair studies are not frequently seen in clinical trials randomizing individual subjects to different intervention groups. However, they have proven to be the design of choice for many investigators embarking on a community intervention trial largely because of the perceived ability of this design to create intervention groups that are comparable at baseline with respect to important prognostic factors, including, for example, community size and geographical area. The relatively small number of communities that can be enrolled in such studies further enhances the attractiveness of pair matching as a method of reducing the probability of substantively important imbalances that may detract from the credibility of the reported results.

Freedman et al. (1990) investigated the gain in efficiency obtained from matching in a community intervention trial. This was done in the context of the COMMIT trial (Gail et al. 1992). Eleven pairs of communities were matched on the basis of several factors expected to be related to the smoking quit rates, such as community size, geographical proximity, and demographic profile. Within each matched pair, one community was allocated at random to the intervention group, with the other acting as the control. It is also interesting to note that this trial was one of the first large-scale community intervention studies to use formal power considerations at the planning stage and, perhaps not coincidentally, to be substantially larger in size than its predecessors.

The gain in efficiency (measured by the sample size required for a given power and effect size) due to matching may be quantified by the factor $G = 1/(1 - \rho_{\mathrm{m}})$,

where ρ_m is the correlation between members of a pair with respect to the outcome variable. This latter quantity is simply the Pearson correlation between outcomes for the intervention and control. Thus if the correlation was 0, there would be no gain in efficiency, whereas a value of 0.5 would reduce the required sample size by 50% if matching were employed. Freedman et al. (1990) showed that matching can lead to considerable gains in statistical precision when it is based on an effective surrogate for outcome. However, since G is simply the ratio of population variances ignoring or accounting for pair-matching, it does not take into account the difference in degrees of freedom for estimating these variances, a factor which is particularly relevant in trials enrolling a small number of communities. For example, in the COMMIT study, there were 11 matched pairs. The degrees of freedom associated with the error from the paired differences in event rates would be only ten, as compared to the 20 degrees of freedom available for an unmatched analysis. This issue was subsequently addressed in detail by Martin et al. (1993), who concluded that for studies having no more than 20 pairs, matching should be used for the purpose of increasing power only if the investigators are confident that ρ_m exceeds 0.20. By considering the practical difficulties that often arise in securing "good" matches, they also concluded more generally that "matching may be overused as a design tool" in community intervention trials.

These considerations suggest that a tempting strategy in practice may be to perform an unmatched analysis of data arising from a matched pair design, particularly when matching is adopted mainly for the purpose of avoiding a "bad" randomization. The effectiveness of such a strategy was investigated by Diehr et al. (1995), who concluded on the basis of an extensive simulation study that breaking the matches can actually result in an increase in power when the number of pairs is less than ten. Thus the loss in precision identified by Martin et al. (1993) in the presence of weak matching correlations can be at least partially regained.

A secondary objective of many community intervention trials is to investigate the effect of individual level risk factors on one or more outcome variables. Focusing on the case of a continuous outcome variable, Donner et al. (2007) showed that the practice of performing an unmatched analysis on data arising from a pair-matched design can lead to bias in the estimated regression coefficient and a corresponding test of significance which is overly liberal. However, for large-scale community intervention trials, which typically recruit a relatively small number of large clusters, such an analysis will generally be both valid and efficient.

10.2.3.2 Limitations of Matched Pairs Designs

Klar and Donner (1997) explored some further limitations of the matched pair design that are more general in nature. These limitations arise largely from the total confounding of the intervention effect with the natural variation that exists between two members of a matched pair. One consequence of such confounding is that it precludes the use of standard methods for estimating the underlying *ICC*, which in turn reduces analytical flexibility. For example, a secondary objective of many studies is to estimate the effect of selected individual level risk factors on one or more outcome variables using regression modeling. However, the calculation of appropriate

standard errors for the regression coefficients obtained from such a model requires a valid estimator of ρ. Thus, although it is possible to perform adjustments for the effect of such risk factors, the task of testing for their independent relationship with outcome is more difficult (Donner et al. 2007). It is difficult to directly model the joint effects of cluster level and individual level risk factors, and the matched pair design frequently does not bring large gains in precision. Klar and Donner (1997) recommended that greater attention should be paid to the possibility of adopting a stratified design, in which two or more clusters are randomized to each intervention group within strata. This design may be particularly attractive when investigators find it challenging to create matched pairs that correspond to unique estimates of risk for each pair. Most importantly, the cluster level replication inherent in this less rigid allocation scheme removes many of the analytical limitations associated with pair-matching, while increasing the degrees of freedom available for estimating error.

Perhaps the most commonly adopted matching factors in large-scale community randomized trials have been cluster size and geographical area (e.g., urban vs. rural). Matching by cluster size is attractive not only because it protects against large imbalances in the number of subjects per intervention group, an efficiency consideration, but also because cluster size may be associated with other important but unaccounted for baseline variables, such as socioeconomic status and access to health-care resources (Lewsey 2004). Matching by categorized levels of baseline versions of the trial outcome rate would also seem attractive. However, results reported by Feng et al. (1999) suggest that if the primary interest is in change from baseline, such matching is not likely to add benefits in power beyond that yielded by an analysis of change scores. This is because including the baseline in the model analysis is as effective as matching for baseline.

10.2.4 Problems with Identifying and Recruiting Patients to Cluster Trials

In the trial conducted by Kinmonth et al. (1998), the subjects were newly diagnosed people with type II diabetes. The doctors who recruited them were the same doctors who were given training in patient-centered care. After the trial, it was discovered that there were more patients diagnosed in the intervention arm than in the control possibly because the doctors were unblind to treatment. However, concealment of allocation is usually regarded as crucial for individually randomized trials, and one of the advantages of randomization is that it ensures that it is impossible to predict which treatment the next potential recruit will get (see also chapter ▸Clinical Epidemiology and Evidence-Based Health Care of this handbook). This advantage is lost for cluster trials where randomization of clusters is usually accomplished at the start of the trial and so concealment is more difficult. In most cases, it is impossible to conceal the identity of the treatment from the patients when they receive it but it is useful to conceal what treatment patients will get until after they are recruited to the trial. In a new trial, currently in the planning stage, an insulin pump is being tested in patients with type II diabetes. The patients are

educated in the use of the pump in groups of size six. The patients are recruited to the trial and asked to give consent to either treatment. When six have been recruited, they are randomized to either the pump therapy or control. In this way the recruiters are ignorant of the treatment the patients will receive. Eldridge et al. (2010) also discussed various options for trying to ensure concealment. These include recruiting clusters and patients before randomization, masking recruiters, or using a standardized recruitment procedure across clusters to try and ensure the procedure was not affected by subsequent treatment.

10.3 Analysis of Cluster Randomized Trials

10.3.1 Cluster Specific Versus Marginal Models

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 the outcome Y. We assume the Z_is have the same distribution for all i. We add a covariate X for the treatment effect, where $X = 1$ for the intervention and $X = 0$ for the control. Suppose the effect of an intervention is to add an amount β_1. We assume that the actual cluster effect is a separate and independent effect to that of the treatment effect. A *cluster specific (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. A suitable model might be

$$E(Y_i|Z_i) = \beta_0 + \beta_1 X + Z_i, \tag{10.5}$$

where $E(Y_i|Z_i)$ is the expected value of the outcome conditional on Z_i. We further assume that $E(Z_i) = 0$ and $\text{var}(Z_i) = \sigma_Z^2$ and the Z_is are independent of the fixed effect X for all i. The distribution of Z_i is generally assumed to be normal, but for binary data, a gamma distribution can be used. In a Bayesian context, other distributions such as a t-distribution can be also used (see Sect. 10.3.4).

Equation 10.5 can be generalized to any outcome variable Y_i (continuous or binary), with expected value μ_i and a generalized link of the form

$$g(\mu_i) = \beta_0 + \beta_1 X + Z_i, \tag{10.6}$$

where the function g is assumed strictly monotone and differentiable.

However, in a cluster randomized trial, everyone in a cluster receives the same treatment, and although a *CS* model can be fitted, the result can be interpreted only theoretically. There is an analogy here with the "counterfactual" argument for causality in epidemiology (see also chapter ▸Basic Concepts of this handbook), where we interpret casual effects as being the difference in outcome from either exposure to a hazard or non-exposure in the same person, even though in practice this is not observed.

An alternative method is to fit a model which looks at the average effect of X over the range of Z. This is the so-called *population averaged (PA)* or *marginal* model. Consider a model where we fit only X and ignore Z_i, so that

$$g(\mu_i) = \beta_0^* + \beta_1^* X. \tag{10.7}$$

Model (10.7) is a *PA* model, that is, we estimate the effect of X on Y as averaged over all the clusters i.

Neuhaus and Jewell (1993) contrasted the approach between cluster specific models and population averaged models by observing that Model (10.7) is simply Model (10.6) with the variable Z omitted. If we assume that the coefficients for X in the two models are related by $\beta_1^* \approx B\beta$, where B is the bias factor, then they show that for a linear model and a log-linear model $B = 1$, and so the interpretation of cluster specific and marginal models is the same. However for a logistic link, $\mu = \text{logit}\,(P(Y = 1|X, Z))$, they showed that $B \approx 1 - \rho$ where ρ is the correlation of the Ys within clusters assuming $\beta_1 = 0$. Since $0 < \rho < 1$, so $0 < B < 1$, and so the general effect of using a population averaged model is to attenuate the regression coefficient toward zero. One can also see that for a logistic link, the greater the variability of the random variable Z_i, and so the greater the intracluster correlation, the greater the attenuation. However, as discussed earlier, the value of ρ in community randomized trials is usually less than 0.01, and this suggests that the bias in assuming a marginal model should not be great.

Since $B = 1$ for a log link, this would suggest that in prospective studies such as clinical trials, it would be advantageous to use a log-linear model which estimates the relative risk rather than a logistic model which yields odds ratios (Campbell 2008). However, experience has shown that in general logistic models are easier to fit and have fewer convergence problems. These can arise with a log-linear model when the fitted values for the risk become greater than 1 or less than 0. This is more likely to happen when the number of events is relatively high.

10.3.2 Standard Methods of Analysis

10.3.2.1 Inflating the Standard Error

For a linear model, we assume the observed outcome y_{ij} is the outcome of the random variable Y_{ij} for the jth subject ($j = 1, \ldots, m_i$) in the ith cluster ($i = 1, \ldots, k$), and it differs from the expected value by a random error. We write

$$Y_{ij} = \beta_0 + \beta_1 X_i + Z_i + \varepsilon_{ij} \tag{10.8}$$

and we assume $E(\varepsilon_{ij}) = E(Z_i) = 0$, ε_{ij} and Z_i are independent, $\text{var}(\varepsilon_{ij}) = \sigma^2$, and $\text{var}(Z_i) = \sigma_Z^2$. In a trial with no other covariates, X_i is an intervention indicator variable (0, 1) which depends only on whether the cluster i is in the intervention

group or not. An *exchangeable* correlation structure is assumed, which effectively means that one can exchange subjects j and j' within a cluster without changing the covariance. This breaks down if subjects are measured more than once (e.g., at baseline and at follow-up) since the correlation of the same subject measured twice will not be the same as the correlation of two different subjects within a cluster. It also means that the *ICC* must be assumed to be the same within each arm of the trial, an assumption which is guaranteed in a randomized trial under the null hypothesis of no intervention effect, but may not be true under the alternative hypothesis that the intervention affects the outcome.

Let $\bar{d}$ be the estimate of the difference in means between the intervention and control group and suppose there are $k/2$ clusters in each group (k assumed even). Then we can show that

$$\text{var}(\bar{d}) = \frac{4\sigma^2}{mk} + \frac{4\sigma_Z^2}{k}. \tag{10.9}$$

The first term in (10.9) is simply the variance that would have been obtained if the data were not clustered. Equation 10.9 can be rewritten as

$$\text{var}(\bar{d}) = 4(\sigma_Z^2 + \sigma^2)VIF/mk,$$

and we can estimate $\sigma_Z^2 + \sigma^2$ by the pooled variance of the outcome variable over groups. Thus a technique originating in sample survey is to simply multiply the variance obtained from ignoring the clustering by the variance inflation factor *VIF*. The design effect given in Sect. 10.1 may be estimated by replacing the *ICC* with its sample estimate (Donner and Klar 2000). A simple test of whether a parameter is zero, known as a modified Wald test, is to divide an estimate of the parameter by its modified standard error which is then compared to the quantile of a standard normal distribution. The authors 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," "robust" or Huber-White estimator (Huber and Ronchetti 2009) which has a long history in econometrics for estimators with continuous data and with heterogeneous variances. The advantage of the robust standard error is that one does not need to estimate the *ICC* separately before conducting the analysis.

10.3.2.2 Summary Measures

A simple method, which is applicable to both binary and continuous data, is the method of summary measures, as popularized by Matthews et al. (1990). For continuous data, one uses the mean of each cluster, and for binary data, one would use the proportion of events (or a transformation such as the logit). This works best when the clusters are all approximately the same size. It gives equal weight to each cluster, irrespective of size, and is a cluster specific method. It has a great deal to recommend since it simply uses the summary statistics for each cluster and is easy

to apply without specialist software. However, one cannot adjust for individual level covariates directly using this approach.

10.3.2.3 Generalized Estimating Equations

The generalized estimating equations (GEE) method, developed by Liang and Zeger (1986) in the context of longitudinal studies, has proved to be very popular for the analysis of data arising from cluster randomized trials. It fits the *PA* model and uses an iteratively reweighting algorithm to estimate the parameters and a robust method (the "sandwich" estimator) for the standard error. It is described in more detail in chapter ▸Generalized Estimating Equations of this handbook. Use of the GEE yields a "shrinkage" estimator which is a compromise between no weighting and weighting by the sample size. It deals with the correlation within clusters by assuming a "working" correlation and then adjusting it according to the data. In cluster trials the choice is between an independent error structure and an exchangeable error structure. An independent error structure is plausible if in fact the intracluster correlation coefficient of the outcome variable is close to zero. An exchangeable error structure means that one can exchange subjects within a cluster and not change the correlation matrix. An exchangeable correlation structure effectively weights each mean by $m_i/(m_i\sigma_z^2 + \sigma^2)$. This weights the means by the sample size m_i when $\sigma_Z^2 = 0$ and gives equal weight when $\sigma^2 = 0$ or when the cluster sizes are all the same. In practice, estimates of the variance components, s_Z^2 and s^2, are used and so s^2 will always be greater than zero which implies the weight will vary unless the outcome were constant.

The use of robust standard errors means that even if a model has an incorrect variance-covariance structure, valid inferences can still be made. For example, one could have a model with an independent error structure and use robust standard errors. This is the same as using the variance inflated method described in Sect. 10.3.2.

GEE is used widely for hypothesis testing and confidence interval construction because it can control for the influence of potential confounders on outcome without the need to specify an underlying distribution for the sample observations. The robust variance estimation relies on between cluster information to assure the validity of the resulting inferences. It is therefore important to be wary of this approach to community intervention trials where the amount of such information tends to be relatively small.

Feng et al. (1996) recommended for continuous data that GEE should not be applied to trials having 20 or fewer clusters. It has been found that using a t-distribution (with a Satterthwaite type correction for the degrees of freedom to allow for unequal variances) and a technique known as the jackknife (Efron and Tibshirani 1998) improves the estimate of the standard error (Mancl and DeRouen 2001). Pan and Wall (2002) proposed replacing the GEE Wald test by approximate t- or F-tests. Although the proposed procedures showed type 1 errors closer to nominal than the usual Wald test, they were shown to be strictly applicable only in clusters of small size. It is therefore clear that more research is needed on the development of adjusted GEE procedures that can be applied to clusters of the size that typically arise in community intervention trials.

10.3.2.4 Random Effect Models

The alternative method of analyzing data from cluster trials is to use a cluster specific model (10.5 and 10.8). We now have to assume distributions for the two error terms. For continuous data the subject level error is assumed normal and for binary data it is assumed binomial. For continuous data the cluster level error is usually assumed normal, and also for binary data, although sometimes a gamma distribution is used. Although this model does not directly reflect the design of a cluster trial since treatment is contrasted to control within the same cluster, as stated earlier it does provide a valid estimate of the treatment effect. These models are also known as "mixed" models (since they contain a mixture of random and fixed effects) or "hierarchical" models since one can think of a hierarchy of clusters and then subjects nested within clusters.

The probability density of an observation from Eq. 10.8 conditional on Z_i is normal with mean $\beta_1 X_i$. Thus $P(y_{ij}|Z_i) = f(y_{ij}|Z_i, \beta, \sigma^2)$ where $f(\cdot)$ is the normal density function. Within a cluster and conditional on Z_i, we assume the observations are independent and so, given observations $y_{i1}, y_{i2}, \ldots, y_{im_i}$

$$P(y_{i1}, y_{i2}, \ldots, y_{im_i}|Z_i) = \prod_{j=1}^{m_i} f(y_{ij}|Z_i, \beta, \sigma^2).$$

This depends on the random variable Z_i, and to find the expected value, we integrate over possible values of Z_i to get

$$P(y_{i1}, y_{i2}, \ldots, y_{im_i}) = \int_{-\infty}^{+\infty} f(Z_i, \sigma_Z^2) \prod_{j=1}^{m_i} f(y_{ij}|Z_i, \beta, \sigma^2) dZ_i.$$

The full likelihood is the product of the above integrals over k clusters

$$L(\beta, \sigma_Z^2, \sigma^2) = \prod_{i=1}^{k} \int_{-\infty}^{+\infty} f(Z_i, \sigma_Z^2) \prod_{j=1}^{m_i} f(y_{ij}|Z_i, \beta, \sigma^2) dZ_i. \tag{10.10}$$

As discussed in Sect. 10.3.1, binary models using different link functions can estimate different population parameters and so deserve special consideration.

Let Y_{ij} (0 or 1) be the jth observation ($j = 1, \ldots, m_i$) in the ith cluster ($i = 1, 2, \ldots, k$). The cluster specific logistic model, following Eq. 10.6, is

$$\text{logit}\,(\pi_{ij}) = \beta_0 + \beta_1 X_i + Z_i, \tag{10.11}$$

where Z_i is the effect of being in cluster i and where $\pi_{ij} = E(Y_{ij}|X_i, Z_i)$. This model can be extended to include individual specific covariates X_{ij}. The random variable Z_i is assumed to be independent of X_i and may be usually assumed to be normally distributed with mean 0 and variance σ_Z^2 although sometimes a gamma

distribution fits the data better. Given the Z_i, the Y_{ij}s are assumed independently distributed with binomial parameter π_{ij}.

The full likelihood L is given by

$$L(\beta, \sigma_z^2) = \prod_{i=1}^{k} \int \prod_{j=1}^{m_i} \pi_{ij}(\beta, Z_i)^{y_{ij}} \{1 - \pi_{ij}(\beta, Z_i)\}^{1-y_{ij}} f(Z_i, \sigma_Z^2) dZ_i. \quad (10.12)$$

Equation 10.10 can be solved directly to maximize the likelihood with respect to the parameters β, σ_Z^2, and σ^2 but not so Eq. 10.12. An early method for binary outcomes and which avoided the integration is a penalized quasi-likelihood approach, using a Laplace method for approximating the integral (Breslow and Clayton 1993). However, this has been replaced by methods which conduct the integration directly using Gaussian quadrature or other numerical methods to obtain estimates of the regression coefficients. Other methods, using iteratively generalized least squares (IGLS), are commonly used for hierarchical models (Goldstein 2002) and are implemented in the package MlWin.

10.3.3 Examples

10.3.3.1 The Analysis of Continuous Data

Table 10.2 gives the results from the DESMOND study (Davies et al. 2008) of the analysis of the outcome HbA1c%, which is treated as a continuous variable. The first row is the result of using a simple t-test on the means of the clusters. This ignores the size of each cluster. The second row uses a robust correction factor for the standard error. The estimate 0.0792 is what one would get from an analysis ignoring clustering, but the standard error is inflated using a "sandwich" estimator (cf. chapter ▸Generalized Estimating Equations of this handbook). The third row uses an exchangeable error structure and shows the effect of "shrinking" the smaller clusters toward the center. The random effects model using maximum likelihood gives nearly the same outcome as the GEE with exchangeable errors, a common finding. One can see that the GEE (independent errors) gives a smaller p-value than the other methods possibly because the independence assumption is implausible.

It is sensible to plot the residuals from the random effect to check for approximate normality. Figure 10.2 shows a plot known as a QQ plot, which plots the residuals against the value that would have been expected from a normal distribution with

Table 10.2 Results of analysis of continuous outcome HbA1c% from DESMOND

	Treatment effect	Std. Err.	z	$P > z$
t-test using means	0.0615	0.1321	0.47	0.64
GEE (independent errors)	0.0792	0.1118	0.71	0.48
GEE (exchangeable errors)	0.0531	0.1098	0.48	0.63
Random effects (max.lik.)	0.0518	0.1100	0.47	0.64

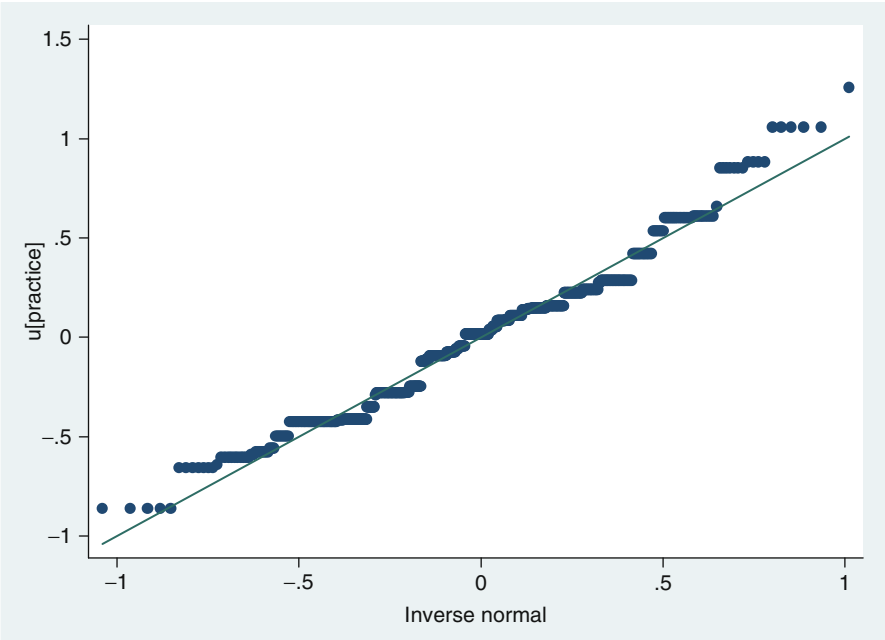

Fig. 10.2 A QQplot of the cluster level residuals from a random effect model from the DESMOND study

the same mean and standard deviation. If the residuals are normally distributed, one would expect this plot to be a straight line. The model fitting procedure means that the residuals are closer to a normal distribution than if we had not estimated the parameters from the data, but the plot is useful for gross departures from expected. The figure suggests that the residuals are plausibly normal. If the plot had been grossly away from normal, one would look for outliers, include potential covariates, and try transformations of the outcome to search for a better fitting model.

10.3.3.2 The Analysis of Binary Data

Table 10.3 shows the outcomes of three analyses from the DESMOND study (Davies et al. 2008) where the outcome is a binary indicator variable (HbA1c% $> 7.5\%$). One can see that the GEE (population averaged) model with exchangeable errors gives a smaller odds ratio (*OR*) than that for independent errors. As with the continuous data analysis of Sect. 10.3.3.1, this is because smaller clusters are shrunk closer to the overall mean in a model with exchangeable errors and it is more likely that small clusters will have larger effects, the effect of which is diminished with exchangeable errors. The *OR* estimated using a random effects model is similar to that for the GEE population averaged model with exchangeable errors since the *ICC* is relatively small (0.061 in this case), but the GEE estimate is very slightly smaller as might be expected. Further discussion of these points has been given by Ukoumunne et al. (2007, 2008).

Table 10.3 Results of analysis of binary outcome for DESMOND using a logistic link

	Odds ratio for treatment	Std. Err.	z	P > z
GEE (independent errors)	1.8318	0.4689	2.36	0.018
GEE (exchangeable errors)	1.8056	0.4568	2.34	0.20
Random effects (adap. quad.)	1.8190	0.5012	2.17	0.30

Tests for the assumptions concerning the residuals with binary data are more difficult to achieve than for continuous data. They are most easily accomplished using Bayesian methods described in the next section. In view of the close agreement between the GEE (exchangeable) and random effects models, we will not pursue this further.

10.3.4 Bayesian Methods

An alternative method to solve Eqs. 10.10 and 10.12 is to use simulation via Markov Chain Monte Carlo (MCMC) algorithms. These are usually associated with a Bayesian analysis, but choice of suitable non-informative priors will yield results similar to the conventional likelihood methods. Spiegelhalter (2001) described methods for the Bayesian analysis of cluster randomized trials with a continuous response. This was extended to a binary outcome by Turner et al. (2001). They used Eq. 10.11 and looked at different prior distributions for σ_Z^2. Since σ_Z^2 is closely related to the *ICC*, they argued that often it is more appropriate to use a prior distribution on the *ICC*, and information for prior distributions for the *ICC* is now becoming available (Gulliford et al. 2005). Turner et al. (2001) experimented with different prior distributions and showed that the estimate of the treatment effect is not entirely robust to the distributional assumptions of the model and suggested caution in using the conventional normality assumption. They showed that the variance components tend to be underestimated when using the non-Bayesian approach. Thompson et al. (2004) and Clark and Bachmann (2009) used Bayesian methods to analyze binary outcomes. They looked at two aspects. Firstly, they looked at rate ratios and rate differences. The latter are particularly important for economic analyses. They showed that use of Bayesian methods facilitated looking at differences in rates. Secondly they looked at the effect of different prior distributions on the outcome. Both sets of authors found that the choice of a prior distribution could have a significant effect on the treatment estimate.

10.3.5 Modeling in Matched Pair Designs

Thompson et al. (1997) replaced standard modeling approaches by techniques borrowed from meta-analysis. Thus an intervention/control pair replaces an individual clinical trial of a meta-analysis. This is essentially equivalent to relying on

between-stratum information to estimate ρ under the assumption of no intervention by stratum interaction. An attractive feature of this is that the forest plot can show which pairs appear to be outliers. However, this approach requires a large number of strata (pairs) to ensure its validity and is therefore not applicable to many community intervention studies. The meta-analysis method as applied to the matched pairs design was extended to binary data by Alexander and Emerson (2005) using a Bayesian approach.

The strict lack of applicability of the t-test to binary outcomes in a matched pair design has led some investigators to alternatively recommend non-parametric approaches, such as Fisher's one sample randomization (permutation) test. Simulations performed by Gail et al. (1996) showed that inferences for matched pair binary data using permutation procedures will have significance levels near nominal under conditions likely to arise in community intervention trials. Essentially the same conclusions were reached by Brookmeyer and Chen (1998) for person-time data arising from matched pair trials. It is useful to note, however, that the one sample permutation test requires a minimum of six pairs to yield a two-sided p-value of less than 0.05, reflecting its relatively weak power. Donner and Donald (1987) showed that a weighted paired t-test based on a logistic transformation of the crude event rates tends to be more powerful than both the permutation test and the standard paired t-test in trials having a small number of strata.

10.3.6 Advice on Methods of Analysis

A method of analysis we have not discussed here is the so-called "fixed" effect method. This involves fitting dummy variables for each of the clusters. This would be relevant if we were particularly interested in the results for particular clusters. However, in general, the clusters are just a source of variation; if the trial were run again, different clusters would be used. Thus treating clusters as fixed incorrectly removes a source of variability, and so the standard errors from this approach will be incorrect.

Heo and Leon (2005) compared different methods of analysis using Eq. 10.12: (1) full likelihood, (2) penalized quasi-likelihood, (3) generalized estimating equations and (4) fixed effects logistic regression. The third method is a marginal method which, following the discussion in Sect. 10.3.1, estimates a different population parameter than the regression coefficient in Eq. 10.10. However, it does not require one to assume a normal distribution for the Z_i. The last method does not take the *ICC* into account and is an invalid method in general for cluster randomized trials. However, if $\sigma_Z^2 = 0$, it may be expected to yield valid tests and efficient estimates.

Heo and Leon (2005) found the full likelihood method and the penalized likelihood methods to be similar and no worse than the fixed effects method even when the within-cluster correlations are zero. As expected, the GEE method gave biased estimates of β_1, the cluster specific parameter for the treatment effect

from Eq. 10.6. They did not investigate the effects of estimating β_1^*, the population averaged parameter from Eq. 10.7. Preisser et al. (2003) showed how to apply a *PA* model to a pretest posttest cross-sectional design, where the assumption of an exchangeable correlation matrix breaks down. They stated that the GEE approach is asymptotically equivalent to the summary measure approach and quoted Mancl and DeRouen (2001) to the effect that using bias-corrected variances can yield valid test sizes even with unequal cluster sizes and with as few as ten clusters.

Ukoumunne et al. (2008) carried out a number of simulations contrasting a cluster level t-test with GEE methods, when the outcome is either the difference in proportions, the risk ratio, or the odds ratio. They found that GEE had little bias in any scale, when the number of clusters per arm was at least ten. In contrast the cluster level t-test only performed reasonably for the difference in proportions.

There are other methods for the analysis of cluster trials such as the use of the bootstrap (Carpenter and Bithell 2002) and methods which fit models in stages. Feng et al. (1996) conducted a comparison of maximum likelihood assuming a normal mixed model, GEE, a bootstrap, and a "4-stage method." The bootstrap used by the authors draws a random sample of size k from the original k clusters. Then one can use ordinary least squares to estimate the βs and repeat a large number of times. The four-stage method is non-iterative, where the first step is to estimate the βs by ordinary least squares and obtain the residuals $e_i = Y_i - X_i\hat{\beta}$. Then the e_is are regressed against the Z_is, leading to estimates of σ^2 and σ_Z^2. One can then use these estimates in a weighted least squares regression of Y_i versus X_i. For small numbers of clusters (<10 per arm) and for nearly balanced data, the bootstrap has been shown to do well, especially if one does not wish to assume normality. For larger numbers of clusters, the maximum likelihood method performed better than GEE.

In practice there is a choice of four main types of analysis: use of summary statistics, generalized estimating equations (GEE), random effects models, and Bayesian random effects models. These are summarized in Fig. 10.3.

As stated earlier, within the GEE method, there is a choice of independent error structure or exchangeable error structure. Within the random effects method, there are a number of ways of fitting models which usually give similar results. For community intervention trials with few large clusters, there is much to recommend a summary measure approach: easy to implement and to understand. If the design includes matched pairs of clusters, then if the number of pairs is less than 10, an analysis ignoring matching is likely to be worthwhile. For "small cluster size" trials with more than 20 clusters per intervention group, the GEE methods using an exchangeable correlation structure are simple and robust. With fewer clusters, one could adopt a random effects model or use a cluster level method. As with any statistical analysis, with few data there is a compromise to be made between the number of assumptions about the data and the power to test hypotheses. With random effects models, it is important to test the assumptions regarding the distribution of the random effects. As stated earlier, this is most easily done using a Bayesian approach, but this requires a degree of expertise and is less commonly done.

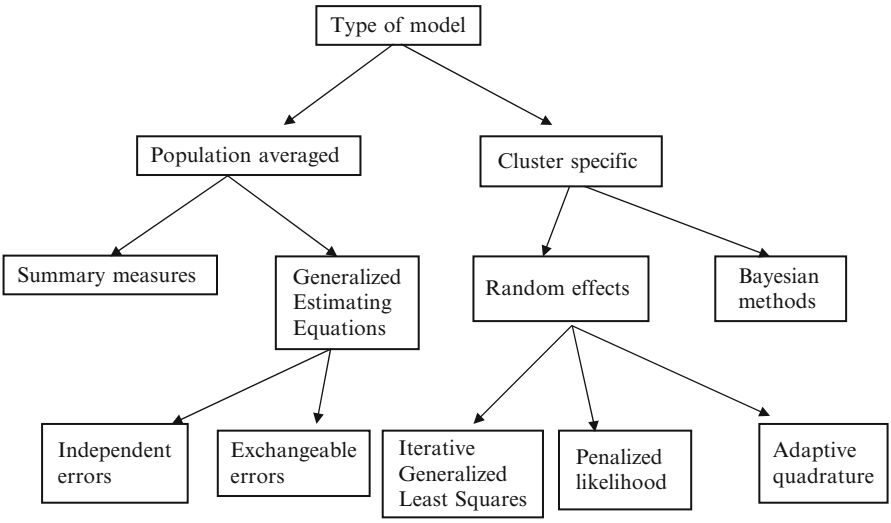

Fig. 10.3 Choices of model and fitting methods

10.4 Other Considerations

10.4.1 CONSORT Statement for Presenting Results from Cluster Randomized Trials

Cluster randomized trials are often poorly reported (Eldridge et al. 2004). The original CONSORT statement was an attempt to improve reporting of individually randomized trials. This statement was subsequently adapted for cluster randomized trials and revised by Campbell et al. (2004). The most important distinctions from the original CONSORT statement are (1) to give a rationale for adopting a cluster design, (2) to describe how the effects of clustering were incorporated into the sample size calculations, (3) to describe how the effects of clustering were incorporated into the analysis, and (4) to describe the flow of both clusters and participants through the trial, from assignment to analysis. Thus, in a primary care trial, for example, one would like to know how any primary care groups were approached, how many agreed to participate, how many were randomized, and how many dropped out during the trial, as well as the characteristics of the patients in the study. The most up-to-date statement is available at www.consort-statement.org.

10.4.2 Clustering by Therapist

Investigators may need to contend with clustering of subjects' responses even for trials using individual randomization. For example, patients may be randomized as they enter a trial by whether or not they will receive a new intervention (say

acupuncture). However, there may be a limited number of acupuncturists and so a single acupuncturist may treat a number of patients. The outcomes will thus be affected not only by whether the patient received acupuncture but by which acupuncturist treats them. The model is

$$Y_{ijk} = \beta_0 + \beta_1 X_i + \gamma Z_k + \varepsilon_{ij},$$

where the subscript k indicates the effect of different acupuncturists and $Z_k = 0$ for the control group. However, it is important to note that the subjects in the control group are not naturally clustered. Issues in the analysis of these trials have been covered by Roberts and Roberts (2002). Lee and Thompson (2005) discussed a Bayesian approach to the analysis of such trials and pointed out how taking account clustering in the analysis can affect the results.

10.4.3 Compliance and Recruitment

We now discuss some potential sources of bias that are peculiar to cluster randomization trials. As we discussed in Sect. 10.2.2.1, if the subjects in a trial are newly diagnosed patients and the intervention is some new approach to treating a disease, then it is possible that practitioners in the intervention arm, being newly educated about this disease, may be more likely to diagnose the disease (Campbell 2000). This may lead to serious problems of selection bias if patients in the intervention arm have less serious disease than patients in the control arm. Trials should be analyzed by what is known as the "intention to treat" (ITT) principle. This means that patients randomized to a particular treatment will be analyzed as if they received that treatment, irrespective of their actual treatment. This is a so-called "pragmatic" approach which attempts to reflect what will happen in practice, when patients will not necessarily comply with treatment. An ITT approach is appropriate when compliance varies over clusters but varying compliance has major implications for any attempt at casual modeling. Loeys et al. (2001) demonstrated how to use standard GEE and random effects models to allow for variable compliance among clusters.

From experience, factors that improve compliance include building a "team spirit" within a cluster, regularly communicating to the patients about the trial, and providing the control group access to the intervention after the trial. For example, in the Hampshire Depression Trial (Thompson et al. 2000), general practitioners (GPs) in the intervention were trained to recognize depression in their patients. GPs in the control group were offered training when the trial was over, and this increased their willingness to stay in the trial.

10.4.4 Software

The selection of the general packages which can fit these models discussed in this chapter is given in Table 10.4. Stata (Statacorp 2009) has simple commands

Table 10.4 Web addresses for software packages

Name	URL
MLwiN	http://www.bristol.ac.uk/cmm/software/mlwin/
R	http://www.r-project.org/
SAS	http://www.sas.com/technologies/analytics/statistics/
Stata	http://www.stata.com/
WinBUGS	http://www.mrc-bsu.cam.ac.uk/bugs/

for applying a cluster robust standard error and for checking the distribution of the residuals. These can also be accomplished in R, but it is not as easy to use. MLwiN enables more than two-level clustering (e.g., by pupil by class by school) or therapist by patient by time within patient. It also can fit models using Markov Chain Monte Carlo methods which enable a Bayesian approach. WinBUGS can be used to analyze trials using Bayesian methodology with prior distributions for the parameters. SAS is particularly flexible for mixed models.

10.5 Conclusions

10.5.1 Review

Cluster randomized trials are an order of magnitude more complicated than ordinary randomized controlled trials. If the risk of contamination is low, then an investigator would be well advised to consider whether an individually randomized trial might be more efficient. However, the last 10 years have seen a flourishing of research into cluster randomized trials and they are now better understood and can be analyzed relatively easily using common software. There is still a need for information about likely values of the intracluster correlation coefficient for common outcomes and clusters so that trials can be planned with more precision. Trial design is still comparatively simple, and research is needed on issues such as group sequential trials where interim analysis can inform the future design of the trial.

10.5.2 Further Reading

The standard text books on cluster (or group) randomized trials are those by Murray (1998); Donner and Klar (2000) and Eldridge and Kerry (2012). A recent book by Hayes and Moulton (2009) emphasizes the use of cluster trials in infectious diseases, particularly in developing countries. There have been a number of reviews of cluster randomized trials. Klar and Donner (2001) and Donner and Klar (2004), following on from their book (Donner and Klar 2000), reviewed developments up until that time and suggested areas of further research. Murray et al. (2004) reviewed methods in public health. Methodological developments for cluster randomized trials have also been reviewed more recently by Campbell et al. (2007).

References

Alexander N, Emerson P (2005) Analysis of incidence rates in cluster-randomized trials of interventions against recurrent infections, with an application to trachoma. Stat Med 24:2637–2647

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:125–134

Brookmeyer R, Chen Y-Q (1998) Person-time analysis of paired community intervention trials when the number of communities is small. Stat Med 17:2121–2132

Campbell MJ (2000) Cluster randomized trials in general (family) practice research. Stat Method Med Res 9:81–94

Campbell MJ (2008) Should we use relative risks or odds ratios in cluster randomized trials with binary outcomes that have high proportions? J Epidemiol Community Health 62(Suppl 1):A24

Campbell MJ, Donner A, Klar N (2007) Cluster randomized trials and *Statistics in Medicine*. Stat Med 26:2–19

Campbell MK, Elbourne DR, Altman DG (2004) CONSORT statement: extension to cluster randomized trials. BMJ 328:707–708

Carpenter J, Bithell J (2002) Bootstrap confidence intervals: when, which what? A practical guide for medical statisticians. Stat Med 19:1141–1164

Clark AB, Bachmann MO (2009) Bayesian methods for analysing cluster randomized trials with count outcome data. Stat Med 29:199–209

Davies MJ, Heller S, Skinner TC, Campbell MJ, Carey ME, Cradock S, Dallosso HM, Daly H, Doherty Y, Eaton S, Fox C, Oliver L, Rantell K, Rayman G, Khunti K (2008) Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomized controlled trial. Br Med J 336:491–495

Diehr P, Martin DC, Koepsell T, Cheadle A (1995) Breaking the matches in a paired t-test for community interventions when the number of pairs is small. Stat Med 14:1491–1504

Donner A, Donald A (1987) Analysis of data arising from a stratified design with the cluster as unit of randomization. Stat Med 6:43–52

Donner A, Klar N (2000) Design and analysis of cluster randomization trials. Arnold, London

Donner A, Klar N (2004) Pitfalls of and controversies in cluster randomized trials. Am J Public Health 94:416–422

Donner A, Birkett N, Buck C (1981) Randomization by cluster: sample size requirements and analysis. Am J Epidemiol 114:906–914

Donner A, Taljaard M, Klar N (2007) The merits of breaking the matches in community intervention studies: a cautionary tale. Stat Med 9:2036–2051

Efron B, Tibshirani R (1998) An introduction to the bootstrap. CRC Press, Boca Raton/New York/Abingdon

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

Eldridge S, Kerry S, Ashby D (2006) Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol 35:1292–1300

Eldridge S, Kerry S, Torgerson D (2010) Bias in identifying and recruiting participants in cluster randomised trials: what can be done? BMJ 340:36–39

Eldridge S, Kerry S, (2012) A practical guide to cluster randomized trials in Health Service Research. Wiley, Chichester

Feldman HA, McKinlay SM (1994) Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med 13:61–78

Feng Z, Grizzle JE (1992) Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. Stat Med 11:1607–1614

Feng Z, McLaran D, Grizzle J (1996) A comparison of statistical methods for clustered data analysis with Gaussian error. Stat Med 15:1793–1806
Feng F, Diehr P, Yasui Y, Evans B, Beresford S, Koepsell TD (1999) Explaining community level-variance in group randomized trials. Stat Med 18:539–556
Flynn TN, Whitley E, Peters TJ (2002) Recruitment strategies in a cluster randomized trial – cost implications. Stat Med 21:397–405
Freedman LS, Green SB, Byar DP (1990) Assessing the gain in efficiency due to matching in a community intervention study. Stat Med 9:943–953
Friede T, Kieser M (2006) Sample size recalculation in internal pilot study designs: a review. Biom J 48:537–555
Gail MH, Byar DP, Pechacek TF, Corle DK (1992) Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). Control Clin Trials 13:6–21
Gail MH, Mark SD, Carroll RJ, Green SB, Pee D (1996) On design considerations and randomization-based inference for community intervention trials. Stat Med 15:1069–1092
Goldstein H (2002) Multilevel statistical models, 3rd edn. Wiley, Chichester
Grosskurth H, Mosha F, Todd J, Mwijarubi E, Klokke A, Senkoro K, Mayaud P, Changalucha J, Nicoll A, Ka-Gina G (1995) Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial. Lancet 346:530–536
Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ (2005) Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. J Clin Epidemiol 58:246–251
Hayes RJ, Moulton LH (2009) Cluster randomized trials. CRC press, Boca Raton/New York/Abingdon
Heo M, Leon AC (2005) Comparison of statistical methods for analysis of clustered binary observations. Stat Med 24:911–923
Hsieh FY (1988) Sample size formulae for intervention studies with the cluster as unit of randomization. Stat Med 8:1195–1201
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, Hoboken
Kerry SM, Bland JM (2001) Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med 20:377–390
Kinmonth AL, Woodcock A, Griffin S, Spiegal N, Campbell MJ (1998) Randomized controlled trial of patient centred care of diabetes in general practice: impact on current well being and future disease risk. Br Med J 317:1202–1208
Klar N, Darlington G (2004) Methods for modelling change in cluster randomization trials. Stat Med 23:2341–2357
Klar N, Donner A (1997) The merits of matching in community intervention trials. Stat Med 16:1753–1764
Klar N, Donner A (2001) Current and future challenges in the design and analysis of cluster randomization trials. Stat Med 20:3729–3740
Lake S, Kamman E, Klar N, Betensky R (2002) Sample size re-estimation in cluster randomization trials. Stat Med 21:1337–1350
Lee KJ, Thompson SG (2005) The use of random effects models to allow for clustering in individually randomized trials. Clin Trials 2:163–173
Lewsey JD (2004) Comparing completely and stratified randomization designs in cluster randomized trials when the stratifying factor is cluster size: a simulation study. Stat Med 23:897–905
Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Loeys T, Vansteelandt S, Goetghbeur E (2001) Accounting for correlation and compliance in cluster randomized trials. Stat Med 20:3753–3767
Machin D, Campbell MJ, Tan S-B, Tan S-H (2008) Statistical tables for the design of clinical studies. Wiley-Blackwell, Chichester
Mancl L, DeRouen TA (2001) A covariance estimator for GEE with improved small sample properties. Biometrics 57:126–134

Martin DC, Diehr P, Perrin EB, Koepsell TD (1993) The effect of matching on the power of randomized community intervention studies. Stat Med 12:329–338
Matthews JNS, Altman DG, Campbell MJ, Royston P (1990) Analysis of serial data using summary measures. Br Med J 300:230–235
Moerbeek M (2006) Power and money in cluster randomized trials: when is it worth measuring a covariate? Stat Med 25:2607–2617
Mills S, Campbell MJ, Waters WE (1986) Public knowledge of AIDS and the DHSS advertisement campaign. Br Med J 293:1089–1090
Murray DM (1998) The design and analysis of group-randomized trials. Oxford University Press, Oxford
Murray DM, Vernell SP, Blitstein JL (2004) Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health 94:423–432
Neuhaus JM, Jewell NP (1993) A geometric approach to assess bias due to omitted covariates in generalized liner models. Biometrika 80:807–815
Nixon RM, Thompson SG (2003) Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials. Stat Med 22:2673–2692
Pan W, Wall MM (2002) Small sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat Med 21:1429–1441
Preisser JJ, Young ML, Zaccaro DJ, Wolfson M (2003) An integrated population average approach to the design, analysis and sample size determination of cluster unit trials. Stat Med 22:1235–1254
Roberts C, Roberts SA (2002) Design and analysis of clinical trials with clustering effects due to treatment. Clin Trials 2:152–162
Soncini JA, Maserejian N, Trachtenberg F, Tavares M, Hayes C (2007) The longevity of amalgam versus compomer/composite restorations in posterior primary and permanent teeth. J Am Dent Assoc 138:763–772
Spiegelhalter DJ (2001) Bayesian methods for cluster randomized trials with continuous response. Stat Med 20:435–452
Statacorp (2009) Stata release 11. Statistical software. Statacorp LP, College Station
Thompson C, Kinmonth AL, Stevens L, Peveler R, Stevens R, Ostler K, Pickering RM, Baker NG, Henson A, Preece J, Cooper D, Campbell MJ (2000) Effects of a clinical practice guideline and practice based education on the detection and outcome of depression in primary care: Hampshire depression project randomised controlled trial. Lancet 355:185–191
Thompson SG, Pyke S, Hardy RJ (1997) The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. Stat Med 16:2063–2980
Thompson SG, Warn DE, Turner RM (2004) Bayesian methods for analysis of binary outcome data in cluster randomized trials on the absolute risk scale. Stat Med 23:389–410
Turner RM, Omar RZ, Thompson SG (2001) Bayesian methods of analysis for cluster randomized trials with binary outcome data. Stat Med 20:453–472
Turner RM, Prevost AT, Thompson SG (2004) Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. Stat Med 23:1195–1214
Ukoumunne OC, Thompson SG (2001) Analysis of cluster randomized trials with repeated cross-sectional binary measurements. Stat Med 20:417–433
Ukoumunne OC, Carlin JB, Gulliford MC (2007) A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. Stat Med 26:3415–3428
Ukoumunne OC, Forbes AB, Carlin JB, Gulliford MC (2008) Comparison of the risk difference, risk ratio and odds ratio scales for quantifying the unadjusted intervention effect in cluster randomized trials. Stat Med 27:5143–5155
Wight DG, Raab GM, Henderson M, Abraham C, Buston K, Hart G, Scott S (2002) Limits of teacher delivered sex education: interim behavioural outcomes from randomized trial. Br Med J 324:1430–1433