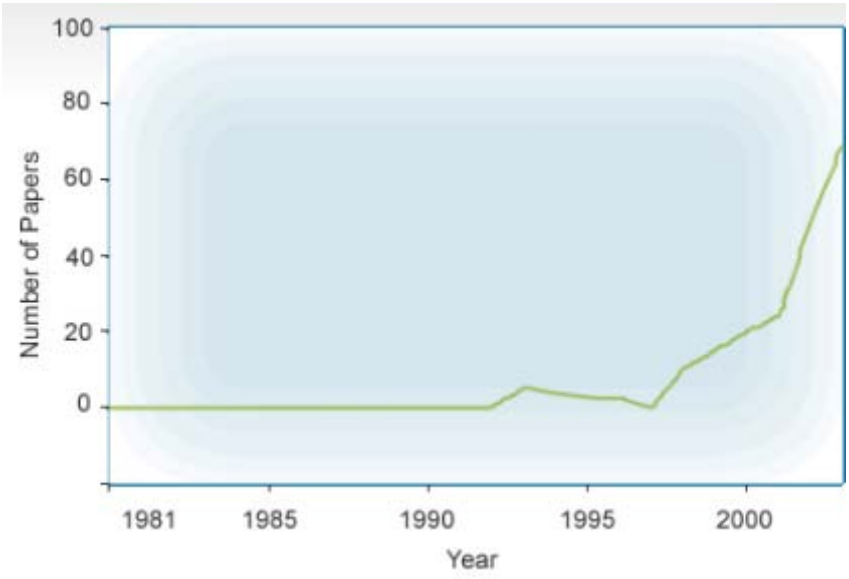
2. Introduction

Cluster randomization trials (CRTs) are experimental studies in which intact social units (clusters), such as families, schools or even entire cities, rather than independent individuals, are randomly allocated to intervention groups. The purpose of this chapter is to provide an introduction to the basic principles that apply to the design and analysis of CRTs, using examples throughout to illustrate the key points.

Although CRTs were seen only infrequently before the mid-1980s, their popularity has increased dramatically over the last 20 years, particularly in the evaluation of innovations in health care. Figure 1, adapted from Bland (2004), shows that the growth in CRTs from the mid-1990s has been particularly rapid. Perhaps not surprisingly, the methodological foundation for CRTs has been much slower to develop, with the first texts dealing exclusively with this design only appearing in the last 10 years (Murray, 1998; Donner and Klar, 2000). A review of more recent developments may be found in Campbell et al., (2007).

Figure 1: Cluster Randomization Trials Published 1981-2003



Source: Bland J.M. (2004) "Cluster randomised trials in the medical literature: Two bibliometric surveys." *BMC Medical Research Methodology*; 4: Figure 1, p: 4.

2. Introduction

In each of the previous examples, the reasons given for randomizing clusters rather than individuals were entirely practical in nature. Thus in Example A the authors stated that it was not "politically feasible" to administer Vitamin A supplementation to some children in a village but not to others. Allocation by village also had the advantage of avoiding the errors that might be introduced if study staff were to administer different capsules to different residents.

The desire to improve compliance was likely a motivating factor in Example B, where adherence to the intervention could be enhanced if all family members were using the same nasal tissues. Logistical considerations were also considerably simplified.

Perhaps the most common reason cited in the literature for adopting a cluster randomization is the risk of experimental contamination.

Experimental contamination would be a serious risk in the breastfeeding trial (Example C) if the same physician was to implement different interventions to women in the same clinic. Similar concerns motivated the randomization of practices in Example D, while randomization at the community level was virtually a necessity in Example E.

2. Introduction

More generally, contamination may be a serious concern in trials of lifestyle modification and counseling that are conducted using individual randomization if there is an opportunity for control group subjects to mingle with experimental group subjects, and therefore to compare or even share the interventions received.

Other reasons to adopt this design arise simply because of administrative convenience. For example, having set up a disease screening program in general practice, it could be logistically awkward and distracting to staff if a formal randomization scheme were implemented within a doctor's office. In this case the CRT design is attractive since it allows physicians and nurses to operate as they would normally on a day to day basis. That is, it allows the intervention to be given in a way that is more consistent with how the intervention would be given in practice. Randomization by practice also removes problems that could arise if some medical professionals have ethical qualms about offering an innovative health care program to only some of their patients.

Medical settings and communities tend to be the most common randomization units seen in the literature. However some CRTs have randomized more unusual clusters, as listed in Table 1.

Table 1: Examples of Unusual Clusters

Cluster	Researcher
Religious institutions	Lasater et al., 1997
Baseball teams	Walsh et al., 1999
Sex establishments	Fontanet et al., 1998
Student pubs	Johnsson and Bergland, 2003
Boy scout troops	Jago et al., 2006
Calendar weeks	Mason et al., 2007
Grocery stores	Hunt et al., 2003

2. Introduction

Exercise 1: Cluster Randomized Trial

Case Study Example:
A trial for the prevention of heart disease in factory workers is being planned. The main purpose of the trial will be to evaluate a strategy in which information concerning coronary risk factors is provided to the workers by a health counselor. The investigators have chosen factories as the unit of randomization.

QUESTION 1 of 3: In the case described, determine what the motivating factor(s) for adopting cluster randomization might have been.

Possible Motivating Factor(s)	
Risk of contamination Choosing to randomize clusters (factories) rather than individuals will minimize the likelihood of subjects in different intervention groups sharing information.	☐
Secure Informed Consent Cluster randomization will minimize problems related to securing informed consent.	☐
Ethics Factory authorities may find it ethically compromising to offer preventive advice to some workers in their factory but not others.	☐
Convenience Providing advice to all eligible factory workers in naturally available groups would be more administratively convenient and less costly than providing such advice on a one-to-one basis.	☐
Reduce Loss to Follow Up Cluster randomization is likely to reduce the anticipated loss to follow-up rate.	☐

QUESTION 2 of 3: Identify what potential problems may arise from cluster randomization.

Potential Problem(s)	
Individual Level Risk Factors Randomization at the cluster level can provide baseline comparability with respect to cluster level risk factors but not with respect to individual level risk factors.	☐
Selection Bias Since the investigators will be obliged to inform half the participating factories that they will not receive the intervention, some factories may decline to accept the assignment of their workers to the control group. This could create problems of selection bias.	☐
Dilution If workers who already have heart disease are not removed from the trial at baseline, the effect of the intervention will be diluted.	☐
Generalizability Factories willing to participate in the trial may be those who are most likely to comply with the intervention. Moreover the "healthy worker effect" must be recognized, where occupational cohorts differ from the general population in their health status. These factors could affect the generalizabilty of the trial results.	☐
Intracluster Correlation Statistical analyses at the cluster (factory) level are fully efficient only if the intracluster correlation coefficient is 1.0.	☐

QUESTION 3 of 3: Identify what statistical implications might arise from methodological challenges.

Statistical Implication(s):	
Sample Size In order to properly estimate the required sample size for this trial, it will be necessary to obtain an a priori estimate of the intracluster correlation coefficient.	☐
Magnitude of Effect With the intervention provided at the cluster level to relatively healthy individuals, its intensity may be less than if applied at the level of the individual worker. This must be taken into account in assessing the magnitude of the likely intervention effect.	☐
Ecological Fallacy It is necessary to address concerns regarding the ecological fallacy in interpreting the trial results.	☐
Stratification Depending on the degree of between-factory heterogeneity and the number of factories to be enrolled in the trial, a decision will need to be made concerning the desirability of at least some stratification by baseline risk factors.	☐

3. Statistical Implications

Statistical Implications of Cluster Randomization

A key feature of cluster randomization trials is that while randomization is at the cluster level, statistical analyses are usually conducted at the individual level.

This discordance between the unit of randomization and the unit of analysis, an issue not usually dealt with in standard statistical texts, creates special methodological challenges at every stage of the trial. These challenges arise essentially because individuals in the same cluster tend to respond more similarly than individuals in different clusters, i.e. the assumption of statistical independence required for the application of standard statistical methods is now violated. Thus the outcome measure is now characterized by two separate sources of variation, one within clusters and the other between clusters.

Within-cluster dependencies may arise from several different sources:

- Subject self-selection is one important factor, as when female patients choose female physicians, or when individuals with respiratory problems choose to live in dry weather communities.
- External factors may also be relevant, as when differences in temperature among nurseries are related to infection rates, or when differences in smoking bylaws influence the success of smoking cessation programs.
- Finally, a variety of internal factors may also lead to between-cluster variation, particularly when individuals respond similarly to an intervention that is provided in a group setting.

3. Statistical Implications

Without extensive empirical data, it is very difficult to distinguish among these potential sources of between-cluster variation. Regardless of the source, however, such variations must be taken into account at all stages of a trial in order to avoid misleading conclusions. Thus failure to do so at the design stage could lead to an underpowered study (caused by an elevated type II error), while failure to do so at the analysis stage could lead to a false declaration of statistical significance (caused by an elevated type I error).

The degree of within-cluster resemblance is typically measured by the magnitude of the intracluster correlation coefficient ρ, which may be defined as the standard Pearson product-moment correlation between any two observations in the same cluster. Provided ρ is non-negative (an assumption generally made in CRTs), this parameter may be equivalently defined as the proportion of overall variance in the trial outcome measure that may be attributed solely to variation between clusters. More formally, we may define $\rho = \sigma^2_B / (\sigma^2_B + \sigma^2_W)$, where σ^2_B represents the variance component between clusters and σ2W represents the variance component within clusters.

Letting $\sigma^2 = \sigma^2_B + \sigma^2_W$ denote the overall variance of the outcome measure, we may write $\sigma^2_W = \sigma^2(1 - \rho)$, which shows how higher values of ρ lead to smaller values of σ^2_W for a fixed value of σ^2, thus enhancing the degree of within-cluster resemblance.

Values of ρ in practice tend to be small and positive. For example, in primary care settings the intracluster correlation coefficient has been found to vary from about 0.01 to 0.05 (Campbell et al., 2000), while in trials randomizing intact communities it may be close to 0.001. Unfortunately these very small values have led some investigators to conclude that their impact on the overall study conclusions is likely to be negligible, and therefore can be ignored in the statistical analysis (e.g., Skinner et al., 2000).

3. Statistical Implications

Dismissing small ρ values as negligible can be seriously misleading, since the impact of clustering depends not only on the magnitude of ρ but also on the sizes of the clusters enrolled in the trial.

Assuming a fixed cluster size m, a more complete measure of this impact is given by the value of the "design effect," given by $[1+(m-1)\rho]$. This expression may also be referred to as the variance inflation factor (VIF), since it measures the percentage increase in the estimated variance of a mean or proportion that can be solely attributed to clustering effects. Consider, for example, a school-based trial where past experience suggests that ρ is likely to be about 0.01. If the investigators decide to randomize schools of size $m = 100$ to each of two intervention groups, the value of VIF will be very close to 2.0, implying that the variances of the resulting means and proportions could be underestimated by as much as 50% if clustering effects are ignored.

The impact of clustering may also be viewed in terms of its impact on the “effective sample size” per cluster, given by $m/[1+(m-1)\rho]=m/VIF$. Thus when ρ achieves its maximum value of 1.0, the total amount of information available from each cluster is no more than that provided by a single individual, while at $\rho=0$ each individual in the trial provides an independent piece of information. More generally, it is clear that the total amount of information available from a CRT enrolling a specified number of subjects is less than that available from an individually randomized trial. This observation is what underlies the classic advice that "randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception and should be discouraged” (Cornfield, 1978). It might also be added that randomization by cluster accompanied by a sample size assessment appropriate to randomization by individual is an exercise in self-deception.

4. Common Designs

Commonly Adopted Cluster Randomization Designs

The reduced effective sample size associated with cluster randomization increases the risk of chance imbalance between intervention groups on prognostically important baseline characteristics. This in turn has strongly influenced investigators to use some form of restricted randomization in the formal allocation scheme. As a result the **matched-pair design**, although seen only infrequently in individually randomized trials, has become very popular for CRTs, particularly when the total number of available clusters is small. This design requires members of a pair (stratum) to be first matched on known risk factors for outcome, with each member then randomized to either the intervention or control group.

In the COMMIT trial (Example E above) the participating communities were matched on several baseline characteristics, including community size, population density, demographic profile, community structure, and geographical proximity. The attraction of such extensive matching is the assurance it provided that the groups compared were well balanced on baseline factors potentially related to smoking cessation. This assurance would seem to be particularly important in a trial involving only 11 matched pairs, since any statistical adjustment for chance imbalance at the data analysis stage would necessarily be limited in scope. The PROBIT trial (Example C) also used this design, matching maternity hospitals with respect to geographic region, number of deliveries per year, and breastfeeding initiation rates at hospital discharge.

4. Common Designs

Two matching factors commonly seen in CRTs are geographic location and some measure of cluster size, as was the case in both examples C and E. Matching on cluster size may have multiple benefits since it:

1. Assures that the total number of individuals in each group is approximately the same (an efficiency consideration); and
2. Controls for the possibility that the number of individuals in a randomization unit reflects existing within-cluster dynamics or other factors that are potentially related to outcome.

For example in trials conducted in developing countries, larger villages may have better health outcomes simply because they are located closer to central health facilities. However other factors may also be strong candidates for matching, depending on the main questions of interest. For example, failure to control for regional differences in socioeconomic status led to interpretational difficulties in a breast cancer screening trial (Alexander et al., 1989).

A design less restrictive than pair-matching is one which allows at least two clusters to be assigned to each stratum of an experimental and control group. Thus it can essentially be regarded as a replication of the completely randomized design in each of several strata. An example is provided by the design employed in the diabetes education trial (Example D), where the stratification variables included training status of the physician and type of contact with the primary care organization. It was also employed in Example B, where families were randomized into one of two treatment groups within each of three strata that were defined by household size.

The Vitamin A trial (Example A) is the only one among those listed above that did not involve some form of restricted randomization. However the very large number of clusters randomized in this landmark study should assure reasonable balance on both known and unknown baseline risk factors; therefore matching or stratification would bring only very limited gains in precision at the expense of added administrative complexity.

5. Pair-Matching

When is Pair-matching Worthwhile

The main attraction of pair-matching on strongly predictive baseline risk factors is the potential increase it brings in statistical efficiency and trial power. Dealing first with the case of a quantitative outcome measure, let $d_j = \bar{Y}_{1j} - \bar{Y}_{2j}$ denote the difference in means for the *j*th pair of clusters, $j = 1,2,...k$. Then the variance of *dj* is given by $2\sigma^2(1-\rho_M)$, where *v* denotes the variance of the outcome measure and the "matching correlation" ρ_M denotes the Pearson product-moment correlation between $\bar{Y}_{1j}$ and $\bar{Y}_{2j}$. This simple result shows that the matched-pair design will always be more powerful than a completely randomized design provided ρ_M is positive. However this result ignores the difference in degrees of freedom used to test the effect of intervention in the two designs. Thus in the completely randomized design the analysis would typically take the form of a two-sample t-test with $2(k-1)$ degrees of freedom, while for a pair-matched design, it would typically take the form of a paired t-test with only $k-1$ degrees of freedom.

This discrepancy will have little impact on power in trials enrolling, say, 30 or more matched pairs. However logistical and cost considerations dictate that many CRTs, particularly those designed to evaluate community-based interventions, are forced to enroll far fewer pairs. The question then arises as to what point the gain in efficiency due to pair-matching on important baseline risk factors outweighs the loss in efficiency resulting from halving the available degrees of freedom. This question was addressed by Martin et al., 1993, who used numerical evaluation to conclude that if the number of pairs *k* is 10 or less, the pair-matched design should only be used if the investigators are confident that the value of ρ_M is at least 0.20. More generally they stated "It is unlikely that effective matching would be possible for small studies. Matching may be overused as a design tool." This point was also made by LaPrelle et al., 1992, who stated that matching on variables poorly related to outcome will "do little but reduce power by shifting the unit of analysis from the individual community to the pair of communities."

5. Pair-Matching

Table 2 lists estimated values of ρM (updated from Donner and Klar, 2000, Table 3.2) for a variety of recently published matched-pair trials. The values reported here clearly indicate that the effectiveness of matching can vary greatly from study to study. For example the HIV prevention trial reported by Grosskurth et al., 1995 generated an unusually high matching correlation of 0.94, while two other community-based trials actually generated negative estimates of ρM. It is also interesting to note that the estimated value of ρM for the COMMIT trial is given by 0.21, barely meeting the criterion given by Martin et al., 1993.

Table 2: Matching Correlations

Source	Unit of Randomization	Number of Pairs	Outcome Variable	Matching Correlation
Stanton & Clemens (1987)	Cluster of Families	25	Childhood Diarhea Rate	0.49
Kidane & Morrow (2002)	Cluster of Villages	12	Childhood of Morality	-0.39
Thompson et al., (1997)	Physician Practice	13	Levels of Coronary Risk Factors	0.13
Ray et al., (1997)	Nursing Home	7	Rate of Recurrent Falling	0.63
Peterson et al.(2002)	School district	20	Prevalence of Smoking	0.34
Haggerty et al., (1994)	Community	9	Childhood Diarrhea Rate	-0.32
Grosskurth et al., (1995)	Community	6	HIV Rate	0.94
The COMMIT Research Group (1995)	Community	11	Smoking Quit Rate	0.21

5. Pair-Matching

Some further insight into these challenges may be gained by realizing that pair-matching is most effective when each of the matched pairs constructed correspond to distinct levels of baseline risk. Although several published trials have enrolled more than 50 matched pairs, the ability to actually construct such a large number of distinct matches is likely to be very challenging in practice. This is because there is often only limited knowledge available on the factors likely to affect outcome. However, even if such knowledge exists, it may not be possible to secure matches for all eligible clusters.

Thus, rather than attempting to construct, for example, 52 matched pairs, it may be more practical to adopt a stratified design with, say, 28 strata enrolling four clusters each. The resulting assignment of two clusters to each of the intervention and control groups within each stratum also has important analytic advantages. These accrue because the assignment of multiple clusters to each stratum allows the intracluster correlation coefficient to be directly computed using routine methods (e.g., Donner and Klar, 2000, Section 6.4). This is not possible in the matched-pair design since the lack of cluster-level replication implies that the natural variation between two matched clusters is totally confounded with the effect of intervention. Without a direct measure of such between-cluster variation, additional assumptions and a fairly large number of matched pairs are needed to estimate ρ (Klar and Donner, 1997).

Exercise 2: Matching Clusters

Question: What factors should a researcher consider when deciding whether to match clusters prior to randomization?

- The likely magnitude of the matching correlation.
- Whether the primary trial outcome is binary or continous.
- The interest of the investigators in evaluating the effect of individual level risk factors on outcome.
- The total number of clusters available to be randomized.
- Whether the intervention is to be offered at the cluster or individual level.
- The relative sizes of the clusters to be randomized.

6. Unit of Inference

Specifying the Unit of Inference

A key feature of cluster randomization trials is that the unit of inference is often at the individual level while randomization is performed at a higher level of subject aggregation. This was the case in the hypertension screening trial reported by Bass et al., 1986, which aimed to evaluate the impact of screening on cardiovascular outcomes in individual patients. Although medical practices were chosen as the unit of randomization, this choice was driven entirely by practical considerations, including administrative convenience and the desire to avoid experimental contamination. Similar considerations applied in the design of the Vitamin A trial described in Example A, where villages were randomized to either the experimental group or a control group. However, studies of Vitamin A supplementation have also been carried out using several other units of randomization, including individuals, households, neighborhoods, and entire communities (West et al., 1991). In each of these trials it was the individual that was the unit of inference.

In some trials the unit of randomization and the unit of inference are both defined at the cluster level, which removes the need to adjust for clustering effects.

Example 2: Cluster-Randomized Trials

Althabe et al., 2004 report on a matched-pair trial aimed at reducing the rate of caesarian section deliveries in Latin American maternity hospitals. The intervention in this trial required the obstetrician to seek a second opinion from a senior colleague before proceeding with the c-section, with outcomes recorded at the hospital level only.

Likewise, Diwan et al., (1995) evaluated a policy of “group detailing” on the prescribing of lipid-lowering drugs in a trial randomizing community health centers. A primary endpoint in this study was the number of appropriately administered prescriptions per month, with the health center serving as the unit of analysis.

In both these trials outcomes on any one individual are not of direct interest. Therefore from the perspective of sample size assessment and choice of analysis, the challenges involved are essentially the same as those that apply to individually randomized trials.

These distinctions imply it is important for investigators to clearly specify the primary unit of inference at the planning stage of their trial. Unfortunately this issue has sometimes been referred to as the "unit of analysis problem" (e.g., Whiting-O'Keefe et al., 1984; Divine et al., 1992). Although intended to emphasize the need for accounting for clustering effects when the unit of analysis is at the individual level, this terminology has sometimes been interpreted to imply that all CRTs should use cluster level analyses. On the contrary, it is the unit of inference that determines the level at which the analysis is conducted.

Exercise 3: Unit of Inference

Read each example below and determine the primary unit of inference. Drag and drop the primary unit of reference into the appropriate example.

Individual | Cluster

Example	
The purpose of the study is to evaluate the effectiveness of different guideline implementation strategies in primary dental care. Participating dental practices were randomized to receive one of four different clinical guidelines related to the management of impacted teeth. The primary outcome was the proportion of patients in a practice whose treatment complied with the guideline.	
The purpose of the study is to evaluate the effectiveness of a cholesterol screening program intended to lower the risk of cardiovascular mortality in a trial randomizing family practices to either the intervention strategy or a control. The influence of a patient's level of education on the success of the intervention is also of interest.	
Investigators wish to evaluate the effectiveness of treated nasal tissues versus placebo tissues in reducing incidence of respiratory illness at 24 weeks. In order to enhance compliance, they elect to randomize families rather than individuals to one of the two intervention groups.	

7. Sample Size Assessment

The usual approach to sample size estimation for cluster randomization trials is to multiply formulas found in standard clinical trial textbooks by an estimate of the variance inflation factor *VIF = [1+(m-1) ρ]*. For example, the number of subjects required to compare two means in a completely randomized design that allocates clusters of size m to each of two groups is given by

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \ (2\sigma^2)[1+(m-1)\rho]}{(\mu_1-\mu_2)^2}$$

where $\mu_1-\mu_2$ denotes the magnitude of difference to be detected, σ^2 denotes the variance of the targeted outcome measure, and $Z_{a/2}, Z_{\beta}$ denote the critical values of the standard normal distribution corresponding to a two-sided significance test with error rate α and power $1-\beta$, respectively. Equivalently, the required number of clusters per group is given by $k=n/m$.

The formula above may also be written as

$$Z_{\beta} = \{km/\{[1+(m-1)\rho]2\sigma^2\}\}^{1/2} |(\mu_1-\mu_2)| - Z_{\alpha/2}$$

which more directly shows the increase on trial power (corresponding to increasing values of Z_{β}) obtained by varying the values of *k* and *m*. This version of the formula makes it clear that while power can be improved indefinitely by increasing the number of clusters randomized *k*, increasing their size m can only increase power to a certain point, as limited by the values of *k* and *ρ*. Indeed, one can show that even if all clusters enrolled are (theoretically) of infinite size, it will be impossible to achieve a power of 80% if the number of randomized clusters is insufficiently large.

7. Sample Size Assessment

Most trials enroll clusters of varying size. It is common in this case to replace m in the previous formula by the mean cluster size $\overline{m}$, which will lead to a slightly underpowered study. However if previous data are available on the distribution of cluster sizes, a more accurate formula may be applied (Eldridge et al., 2006). Let $cv=S_m/\overline{m}$ denote the coefficient of variation characterizing this distribution, where S_m is the standard deviation of the cluster sizes. Then VIF may be replaced in the formula above by
$VIF_A=1 + [(cv^2 +1)\overline{m}-1]\rho$. This adjustment has been shown to have greatest impact when the number of clusters is small and/or the value of ρ is high (Guittet et al., 2006).

Example 3: Calculating Sample Size Assessment

Consider a family randomized trial designed to evaluate the efficacy of a dietary intervention in lowering blood pressure. Data from previous trials performed in a similar population indicate that the intracluster correlation coefficient with respect to diastolic blood pressure may be taken as 0.20, while the mean and standard deviation of the corresponding family size distribution can be reasonably estimated as 2.2 and 0.65, respectively (cv=0.30). Previous experience also indicates that the between-subject standard deviation of diastolic blood pressure is approximately 10.0.

Assuming it is of interest to detect a mean difference of 4mm Hg with 80% power at the two-sided 5% level, the value of VIF_A may be obtained as 1+[(0.302+1)2.2-1]0.2 =1.28 and the number of subjects required in each of two groups by

$n= \{(1.96+0.84)2\ 2(102)/42\}\{1+[(0.302+1)2.2\ -1])0.2\}=(98.75)(1.28)=127$ or about 64 families per group.

7. Sample Size Assessment

Methods of sample size estimation that may be used to compare other population parameters of interest, such as proportions and incidence rates, follow the same general principles, as discussed by (Donner and Klar 2000, chapter 5). For example, to compare two proportions P_1 and P_2, the required sample size may be obtained by replacing $2\sigma^2$ in the formula for comparing two means by $P_1(1-P_1)+P_2(1-P_2)$ and $\mu_1-\mu_2$ by P_1-P_2. However, despite the wide availability of such methods, several reviews of cluster randomization trials performed over the last 20 years show that far fewer than 50% of such trials report their actual use in practice (Donner et al., 1990; Simpson et al., 1995; Smith et al., 1997; Varnell et al., 2004; Murray et al., 2008). However an exception to this discouraging trend might be emerging in the field of primary care, where a recent review by Eldridge et al., 2008 found that 62% of trials reviewed accounted for clustering effects in the sample size calculations, a vast improvement compared to results seen in previous reviews.

Values of ρ required for sample size estimation are usually obtained from trials involving the same endpoint and a similar unit of randomization. Fortunately investigators now tend to report this value fairly frequently. Indeed some researchers have now reported estimates of ρ obtained over a range of studies in a particular research area (e.g., Campbell et al., 2000; Murray et al., 2000; Argarwal et al., 2005; Parker et al., 2005; Gulliford et al., 2005). However, the difficulty remains that many such estimates are based on a relatively small number of clusters, and are consequently subject to considerable uncertainty. Therefore it is usually advisable for investigators to perform a sensitivity analysis in which the impact of different values of ρ on the required size of sample can be carefully explored.

For the matched-pair design, the simplest approach to sample size estimation would be to:

1. Compute the required number of subjects using standard formulas for the completely randomized design; and
2. Multiply the result by the factor $1-\rho_M$, where ρ_M is an estimate of the likely size of matching correlation.

If such an estimate is not available from previous data, a conservative approach would be to assume that matching is ineffective, i.e. to use the completely randomized formula directly.

9. Cluster Level Replication

Some of the earlier community intervention trials addressing cardiovascular risk factors enrolled only two clusters, one allocated to the experimental intervention group and the other to a control (e.g., Turpeinen et al., 1979), with justification resting largely on cost and logistical considerations. This two-cluster design, which can still be seen in the literature today, may be very useful for exploratory purposes as a prelude to a more definitive trial that adopts formal power considerations.

Yet the ability to secure a valid estimate of ρ depends on the ability to obtain an accurate estimate of such variation, and hence the design itself is invalid. It is only under the unlikely (and untestable) assumption that ρ is zero that a valid test of the intervention effect can be conducted. Otherwise the results are subject to the same problems of interpretation that would arise in an individually randomized trial that assigns exactly one patient to each of two treatments. Taking a series of repeated measures in each of the two clusters improves the value of this design as an exploratory tool, but does not remove the basic problem.

The main issue here is not power but rather the threat to trial validity, since the effect of intervention is inevitably confounded with the natural variation that exists between the two clusters.

This is not to say that trials randomizing, say, three or four clusters to each group should be encouraged, since, although technically valid, they will almost surely lack the ability to detect important intervention effects. It is only by adopting a formal probabilistic approach to sample size estimation (discussed earlier) that this problem can be avoided.

10. CRTs and Informed Consent

Cluster Randomization Trials and the Need to Obtain Informed Consent

The need to secure informed consent is well established in individually randomized trials, as governed by principles dating back to the time of Hippocrates.

Norms regarding the need to obtain informed consent in trials randomizing intact social units have yet to receive full acceptance.

This is partly due to the great diversity that can be found in the size and nature of units that can be randomized (e.g., families, hospitals, cities) and partly because consent can be obtained, at least theoretically, at multiple levels.

- At the first level, consent is typically obtained from a “gate-keeper” such as a physician, mayor, or school principal, to allow their cluster to be randomized.
- At the community level this approach could be implemented in accordance with guidelines published by the World Health Organization and the Council for International Organizations of Medical Sciences.

These CIOMS guidelines suggest that the gate-keeper should sign a consent form clearly setting out the steps that will be taken for safeguarding the interests of the study participants. The implication here is that such a contract will adequately substitute for the need to obtain informed consent on an individual basis, which from a practical perspective will often be extremely difficult, if not impossible. This approach is also consistent with what has been recommended for “cluster-cluster” trials, in which the intervention is administered to the entire cluster, as in the case of a media message, rather than on a one-to-one basis to individual subjects (Edwards et al., 1999). In the latter case (“cluster-individual” trials) both the need and ability to obtain informed consent at the individual level are arguably much greater than when the intervention is “indivisible” at this level. However if the parent cluster has already been randomized in a cluster-individual trial, the limitation remains that one can only ask an individual subject at this stage to consent to an intervention that has been previously assigned.

Given the complexity of the ethical issues raised by this design, it would seem reasonable, at least as a first step, for researchers to more thoroughly report the steps taken to obtain informed consent in their own trial. However an ultimate goal would be to develop broadly acceptable norms that can be applied to a range of ethical issues that have yet to be adequately explored in the context of CRTs. Further discussion may be found in Klar and Donner, 2007a.

11. Cluster vs. Individual Level Analysis

Cluster Level versus Individual Level Analysis

This chapter has discussed the importance of identifying the unit of inference at an early stage of a trial, since this choice plays an important role in determining the unit of analysis. Thus when inferences are directed at the cluster level, as in the trial reported by Althabe et al., 2004, analyses are also invariably conducted at the cluster level.

But in the more frequently arising case where the unit of inference is the individual, analyses can be conducted at either level. The simplest approach in this case would be to collapse the data in each cluster and then to construct a relevant summary measure, such as a mean, slope, or other cluster level statistic. This essentially removes the need to adjust for clustering effects, since randomization assures that the resulting summary measures are statistically independent. It is also interesting to note that in the case of a quantitative outcome and a fixed cluster size a cluster level analysis is fully as efficient as an individual level analysis (e.g., Klar and Donner, 2007b). This can be most easily seen by verifying that an analysis of variance performed on the individual subject responses is algebraically equivalent to a two-sample t-test performed on the cluster means. Thus the statement sometimes seen in the literature which characterizes a cluster level analysis as fully efficient only when $\rho = 1$ is incorrect.

However for variable sized clusters, an analysis at the cluster level that is not properly weighted to take into account the intracluster correlation as well as the cluster sizes will indeed be less efficient than an individual level analysis that takes into account both these factors. Nevertheless the relative simplicity of a cluster level analysis still remains an advantage, albeit with some loss of efficiency and an inability to adjust for individual level risk factors.

12. Perils of Subsampling

The Perils of Cluster Subsampling

As alluded to earlier in this chapter, **increasing the number of clusters enrolled in a trial has a greater impact on statistical power than increasing the size of the clusters randomized**. Moreover the benefit that may be obtained from increasing the number of participants per cluster is inversely proportional to the value of the intracluster correlation coefficient ρ, with the largest gains in power achieved when the number of participants sampled (subsample size) increases from 1 to $1/\rho$ (Donner and Klar, 2004). Thus for trials randomizing entire communities, where values of ρ may be as low as 0.001, very little increase in power will be obtained by sampling more than 1000 subjects from each community. On the other hand, if ρ is about 0.01, as in school-based trials or trials randomizing medical practices, any gain in power diminishes rapidly after 100 students per cluster are enrolled.

Nonetheless, it is not uncommon for an investigator to administer an intervention to all members of a cluster even though the resulting gains in statistical power are minimal. This is often because the extra costs and logistical difficulties involved may not be considered onerous. However in some cases, such as in the COMMIT trial (Example E), it may be felt that the entire cluster (community), not just those individuals directly impacted, might benefit as a result of the synergy and interaction that takes place among cluster members. Some investigators might also have ethical qualms about delivering a new intervention to some but not all members of a cluster.

Exercise 5: Uses of Subsampling

Review the examples and select whether use of subsampling was appropriate or not.

Appropriate	Example	Inappropriate
✓	1. A trial is being planned in which a information package is to be mailed out to individuals in the intervention communities. Since the average population of each community is in excess of 1000, the investigators conclude that selecting a random subsample of individuals within each community will lead to a negligible loss in power.	✗
✓	2. An investigator elects to select a random subsample of subjects from each cluster to be enrolled in a trial. The size of the subsample to be selected from a given cluster is to be proportional to the size of that cluster since this should maximize the power for detecting a significant intervention effect.	✗
✓	3. To avoid the risk of contamination in a community intervention trial in which the communities are geographically proximate, it is suggested that subsampling should be restricted to the geographic center of each community.	✗
✓	4. In a hospital randomized trial of a new approach for managing palliative care , patients are recruited by physicians or their staff prospectively over time after randomization.	✗

12. Perils of Subsampling

The protection that randomization offers CRTs that adopt a subsampling strategy can only be assured if the subjects selected can be regarded as a random sample of all cluster members.

Detect and Treat Bias: in many trials subsampling is done in rather opportunistic fashion, as when staff personnel attempt over time to identify eligible patients in a trial randomizing medical practices. Since such recruitment is almost inevitably done by staff who are unblinded to intervention status, a serious risk of selection bias may arise if personnel in the experimental arm of the trial are more motivated, enthusiastic, or better trained in their recruitment efforts than staff in the control arm (Torgerson, 2001; Farrin et al., 2005). An indication that such bias has occurred, sometimes referred to as 'detect and treat bias,' would be that a much larger number of patients have been recruited in the experimental group than in the control group.

Of greater concern, however, is the possibility that the characteristics of the patients in the two groups may differ, as, for example, when sicker patients have been easier to identify in the experimental group due to greater diligence in recruiting them. In this case, all the benefits of randomization will be subverted.

- Avoiding bias: in some trials, such as those randomizing worksites, all eligible cluster members may be identified prior to randomization, which is the most straigthtfoward way of avoiding this difficulty. Otherwise it may be advisable for someone not directly connected to the trial to take primary responsibility for subject recruitment, ideally in blinded fashion.

  Motivation of subjects: in the previous example, the selection bias due to differential subsampling can be largely attributable to greater diligence on the part of the investigator team in seeking out eligible subjects. However selection bias may also arise because subjects in one of the two groups are more motivated to participate than subjects in the other group, resulting in differential levels of consent. For example consider a school-based trial in which the aim of the intervention is to reduce the dating violence experienced by high school students. It may seem reasonable to

restrict recruitment in this study to students who have dating experience. However recruitment efforts in control schools might be less intensive or adept than in intervention schools because control group teachers seeking consent may be less familiar with the details of the administered program. As a consequence, control subjects who ultimately do provide consent may be particularly motivated to participate, perhaps because of their greater sensitivity to issues surrounding dating violence. This in turn may lead to evaluation bias if the characteristics of students recruited in the two groups differ on factors (some easily measurable and others less so) that are related to the trial outcome.

- Avoiding bias: therefore a preferred approach in this case would be to conduct an intent-to-treat analysis in which all students in a school are requested to report whether they have experienced dating violence. A secondary analysis could then be performed which includes only students who have dating experience. If the overall conclusions from the two analyses do not agree, then the intent-to-treat analysis must be recognized as the only one free from selection bias.

13. Other Perils

Is fear of contamination in individually randomized trials overrated?

It was mentioned in the beginning of this chapter that the most common reason cited in the literature for adopting a cluster randomization design is

the risk of experimental contamination that might arise under individual randomization.

We now examine this issue in more detail. Suppose for illustrative purposes that an extreme form of contamination occurs under individual randomization where a proportion *R* of control group subjects "cross over" and actually take up the experimental invention. Then if we assume that these subjects will experience the same event rate as the experimental group subjects, the difference in event rates that can be detected will be reduced by a factor of 1-*R*.

For example, suppose the investigators enroll enough subjects in the trial to detect a 20 percentage point difference in event rates at a specified probability level. If a proportion *R*=0.30 of control group subjects are now expected to assume the same event rate as subjects in the experimental group, the trial must be redesigned to detect a difference of 0.70(20)=14 percentage points. However to detect this smaller effect size it can be easily shown (Torgerson, 2001) that the original sample size must be inflated under individual randomization by a factor 1/(1-*R*)2 =1/(.70)2=2.04. The investigators may believe that this contamination effect can be avoided by alternatively choosing a cluster randomization design. However the variance inflation factor associated with clustering may be much larger than this.

- Adjusting the trial size by taking into account the anticipated contamination effect may well be the preferred option, at least in terms of required sample size.
- Nonetheless it must also be recognized that under individual randomization the effect size estimated by the trial data will be attenuated by the resulting contamination.
- Thus if the main aim of the investigators is to estimate the "uncontaminated" effect size, the cluster randomization design may still be preferable.

13. Other Perils

Clustering effects in individually randomized trials

The need for accounting for within-cluster dependencies in testing the effect of intervention in trials randomizing intact social units is now widely recognized. It is less recognized that clustering effects may also arise in trials which are individually randomized but in which the interventions are 'cluster-administered." This would include, for example, interventions involving group therapy or counseling as well as those that could be delivered individually, but may be delivered on a group basis for practical reasons, as in the case of exercise classes. The source of these effects could well be similar to those arising in cluster randomization trials, such interactions among cluster members or the effect of having a common leader.

- Failure to account for them may similarly lead to underpowered comparisons at the design stage of the study or an elevated type I error at the analysis stage. Further discussion and practical recommendations may be found in Baldwin et al., 2005 and Roberts and Roberts, 2005.

Misconceptions concerning the ecological fallacy

The ecological fallacy (Morgenstern, 1998) is well-known as a possible source of misinterpretation in epidemiological studies when correlations calculated at the cluster level are also assumed to apply at the individual level. For example, consider the ecological correlation between the percentage of individuals in a community using sunscreen lotion (x) and the percentage of elderly individuals residing in that community (y). It is well-known that the existence of a positive correlation does not imply that elderly individuals tend to use more sunscreen lotion than younger individuals; rather it could be that younger people in communities with predominantly elderly individuals are the ones using sunscreen lotion.

The ecological fallacy arises here because communities usually contain both types of individuals, those who use sunscreen and those who are elderly. However it does not apply when standard analyses are used to evaluate the effect of intervention in a CRT. This can be seen by supposing that the variable y in the example above is replaced by an indicator variable z representing treatment assignment (1=intervention, 0=control), where the aim of the intervention is to promote the use of sunscreen protection.

- In accordance with the well-known “intent-to-treat” principle, every individual in a community is counted at the analysis stage in the treatment group to which they were assigned, thus assuring that communities will be homogeneous with respect to the variable *z*.
- More generally, and contrary to what has sometimes been implied in the literature (e.g., Kreft, 1998), concern for the ecological fallacy in this case is often misplaced.

14. Analyses at the Individual Level

Incorporating Clustering Effects into Standard Statistical Analyses

Earlier, this chapter reviewed how the impact of clustering on sample size requirements can be accounted for by incorporating the value of the "variance inflation factor" (VIF) into standard sample size formulas. A similar approach may be taken to adjusting for cluster effects when analyses are conducted at the individual level, with sample estimates of the VIF now incorporated into standard test statistics. Attention in this section will be focused mainly on binary outcome data, which tend to arise more frequently in cluster randomization trials than continuous, count and time-to-event outcomes. Detailed discussion of statistical methods that can be applied to a variety of outcome variables arising in CRTs may be found in Donner and Klar, 2000, Chapters 6-8.

Example 4: Analyses at the Individual Level

To illustrate one such approach, consider a trial evaluating the effect of tailored general practice guidelines on the proportion of patients with benign prostatic hyperplasia (BPH) that remained under specialist care at 12 months post-randomization (Mollison et al., 2000). Of main interest here is a comparison of event rates observed on 150 patients contributed by 23 experimental group practices to event rates observed on 142 patients contributed by 26 control group practices. Median cluster sizes in this completely randomized trial were 6 and 3.5 in the experimental and control groups, respectively, a difference that can reasonably be attributed to chance.

66 (44%) patients in the experimental group were still under specialist care at 12 months as compared to 77 (54.2%) patients in the control group. Application of the standard Pearson chi-square test with one degree of freedom to these data yields x^2_p =3.05(p=.08), indicating a difference that is statistically significant at the 10% level. However this test fails to account for the similarity of responses (clustering) among patients belonging to the same practice, and therefore overstates the true level of significance. We therefore compute the "adjusted chi-square statistic" x^2_A, obtained by dividing x^2_p by an appropriate estimate of VIF (Donner and Klar, 1994). Application of this procedure, based on an estimated value of ρ given by 0.077, yields x^2_A =1.684(p=.19, one degree of freedom), a result no longer statistically significant at any conventional level. Algebraic formulas for all results presented in this example are given in the Appendix.

14. Analyses at the Individual Level

It is important to note that this procedure does not require that the value of ρ is constant across all pairs of observations that may be constructed within clusters ("the common correlation" assumption). Only the weaker assumption that the average value of ρ remains constant across clusters is required to ensure the validity of x^2_A.

The previous example is typical of how standard test statistics can be extended in straightforward fashion to take into account clustering effects. A similar extension of the well-known Mantel-Haenszel (MH) statistic for combining several 2X2 contingency tables (Mantel and Haenszel, 1959) can be applied to test the effect of intervention in a stratified cluster randomization design (Donner, 1998). Analogous to the relationship between x^2_A and x^2_P, the adjusted Mantel-Haenszel statistic (MHA) reduces to MH when divided by an appropriate estimate of the VIF.

It was mentioned earlier that the confounding of the effect of intervention with the natural variation between two clusters in a matched pair precludes the direct computation of the relevant intracluster correlation coefficient. Thus matched-pair designs are invariably analyzed using cluster level analyses, such as the standard paired t-test as applied to the observed differences in cluster-level event rates. Although this procedure is very robust to departures from the underlying assumptions of normality and homogeneity of variance (see Donner and Klar, 1996), some investigators have chosen to avoid the need to make this assumption by applying a nonparametric analogue of the t-test. For example, in the COMMIT trial, the investigators applied a one-sample randomization test (also known as a permutation test) to compare smoking quit rates in the eleven matched pairs of communities.

14. Analyses at the Individual Level

Covariate Adjustment

Random assignment of clusters assures that baseline variables measured at both the cluster and individual levels should be reasonably well balanced. Yet chance imbalance at either or both levels can still arise, particularly when the number of clusters is small.

If imbalance regarded as substantively important arises on baseline variables that are highly predictive of outcome, they must be controlled for either at the design stage through matching or stratification or, alternatively, at the analysis stage. A detailed discussion of methods for covariate adjustment in cluster randomization trials is beyond the scope of this chapter. However for dichotomous outcomes they largely take the form of extensions of multiple logistic regression, a well-known multivariable procedure widely used in the analysis of data arising from individually randomized clinical trials.

- The extension adopted most frequently is known as **generalized estimating equations** (GEE), a procedure which allows adjustment for the joint effects of cluster level and individual level risk factors without the requirement of parametric assumptions. This procedure, developed by Liang and Zeger (1986), is readily available in standard software packages such as SAS and STATA. Unfortunately its validity can only be assured if the number of clusters randomized is fairly large, at least 10 or more in each group (e.g., Feng et al., 1996; Pan and Wall, 2002).
- Another widely used extension of logistic regression assumes that the logit transformation of an event rate *p*, defined as *log[p/(1-p)]*, follows a normal distribution, with the method of maximum likelihood used for parameter estimation. However this approach also requires a reasonably large number of clusters to ensure its validity.

A highly accessible discussion of the advantages and disadvantages of these two different extensions of logistic regression is given by Bellamy et al., 2000.

15. Interim Analyses

The Role of Interim Analyses

Interim analyses are now a standard feature of individually randomized trials, particularly those with long-term follow-up and life-threatening outcomes. Although such analyses may have several objectives, the primary one is usually based on the need to detect unexpected differences in treatment effectiveness that may warrant early termination of subject accrual and follow-up.

There is no reason in principle that these factors should fail to apply to trials randomizing intact social units rather than individuals, as in, for example, nutritional supplementation trials having subject morbidity and mortality as the primary response variables. Yet formally planned interim analyses have not tended to play an important role in such trials.

A number of reasons may be responsible for this, including:

- The relatively long lag time needed for an intervention to “settle in”;
- The perception that the intervention in question is fairly benign, as in the case of lifestyle modification or behavioral trials; or
- The likely belief that the assumptions underlying the stopping rules most frequently adopted for individually randomized trials, such as that developed by O'Brien and Fleming, 1979, may not hold in trials randomizing clusters.

But for trials in which cluster accrual occurs gradually over time it has now been shown (Zou et al., 2005) that such rules may in fact be safely applied under very general conditions. Of course, as in the case of individually randomized trials, it is vitally important that the treatment-related results be transmitted only to members of an independent data monitoring committee, and otherwise kept confidential.

A secondary aim of an interim analysis may be to reassess the values of some of the parameters used to estimate the required trial size. In cluster randomization trials, such parameters may

not only include the standard deviation (for continuous outcomes) or the event rate in the control group (for dichotomous outcomes), but also the intracluster correlation coefficient. An example of sample size re-assessment for a cluster randomization trial is provided by Lake et al., 2002.

16. Cohort vs. Cross-sectional Designs

The earlier discussion on analysis at the individual level is most pertinent to cohort designs, where each individual in the study is followed up over time. However in studies enrolling very large clusters, such as entire communities, such detailed follow-up may not be possible. Considerable discussion has therefore arisen in the community intervention trial literature as to the relative advantages of this design to a cross-sectional design, in which different groups of individuals are independently sampled and assessed at each of several time periods.

It is acknowledged that the cohort design is theoretically more powerful from a statistical perspective, since it allows an analysis that controls for individual baseline values, thus allowing the effect of intervention to be estimated with more precision.

However, as shown by Feldman and McKinlay (1994), **this advantage must be weighed against the risk of loss to follow-up that arises in any longitudinal study**. The “worst-case scenario” arises when the loss to follow-up is differential across intervention groups, since then the final estimate of intervention effect may be subject to substantial bias.

Even when subject attrition is unrelated to treatment assignment, a large loss to follow-up rate may result in reduced efficiency relative to a cross-sectional design. Other disadvantages of the cohort design, as reviewed by Atienza and King (2002), include:

- A loss of representativeness of the target population related to the aging of the cohort; and
- "Learning effects” that may result from repeated assessments on the same individual.

These considerations suggest that a cohort design is most effective when:

- Participating clusters are of relatively small size,
- The study population is relatively stable and compliant; and
- Follow-up times are not lengthy.

It follows that for studies enrolling large communities, where complete follow-up is rarely feasible, cross-sectional designs have often been preferred, as in the early trials of

cardiovascular health referred to in this chapter.They also may be the inevitable choice for any intervention that is evaluated at the cluster level only.

To avoid the analytic limitations of cross-sectional designs, an approach adopted by some investigators has been to augment this design by subsampling a cohort consisting of a relatively small number of subjects in each community. For example the COMMIT investigators used randomly selected cohorts of heavy and light-to-moderate smokers, respectively, as one means of evaluating the effect of their community-based smoking cessation intervention.

From a conceptual perspective, the choice of design must also be considered in light of how the primary question of scientific interest is posed. Thus if interest focuses mainly on change at the broader community level, cross-sectional designs may be the more natural choice while cohort designs may be more natural if change at the individual level is of most interest. Methods of analysis that are particularly suited to cross-sectional designs have been discussed by Nixon and Thompson (2003) and Ukoumunne and Thompson (2001) for the case of binary outcomes, and by Koepsell et al., (1991) and Murray (2001) for the case of continuous outcomes.

17. Reporting

Reporting of Cluster Randomization Trials

The well-known CONSORT statement for individually randomized trials (Begg et al., 1996; Moher et al., 2001; Altman et al., 2001) has now been extended to cluster randomized trials (Campbell et al., 2004). The principle features of this extension include recommendations to:

- Provide the rationale for adopting a cluster design;
- Specify how the effects of clustering were incorporated into the sample size calculation and the statistical analysis; and
- Present a chart showing the flow of both clusters and individuals through the trial.

An earlier set of guidelines were provided by Donner and Klar, 2000, Chapter 9. Aside from reporting standards that are unique to CRTs, there are some that have become routinely accepted for individually randomized trials, but now need to be reconsidered. This includes the presentation of baseline characteristics, which for CRTs should be provided separately for cluster level characteristics (e.g., geographic area, cluster size) and individual level characteristics (e.g., age, gender). The presentation of baseline cluster level characteristics is straightforward, since the clusters assigned to each group are independently distributed.

Some special caution is required when comparing individual level baseline characteristics.

- Although it is now recognized that the use of significance tests for this purpose is always a logically flawed procedure (e.g., Senn, 1994), this practice can be particularly misleading when applied to CRTs. This is because the test procedures typically used, such as t-tests and chi-square tests, may fail to account for the clustering effects that apply at baseline as well as at outcome. The resulting *p*-values will be biased downwards, potentially leading to an ill-advised decision to adjust for the characteristic (covariate) in question in the statistical analysis.
- Standard deviations for continuous variables that are used for descriptive purposes will also be biased downwards by clustering effects, but only slightly unless the overall sample size is small and the intracluster correlation coefficient is large (White et al., 2005), conditions unlikely to apply in most CRTs.

- Finally it must be recognized that the effective sample size for the variables involved is no longer the number of individuals *n* per treatment group but rather *n/VIF*. Failure to recognize this makes it difficult to accurately compare the amount of information provided by different trials.

18. Appendix

Let m_{ij} denote the size of the jth cluster assigned to the ith group, i=1,2; j=1,2...k, with $M_i = \sum_{j=1}^{k} m_{ij}$ denoting the total number of subjects in group i, and $\hat{P}_i$ denoting the corresponding value of the overall event rate in this group. Then the standard Pearson chi-square statistic with one degree of freedom may be written as

$$x_P^2 = \frac{\sum_{i=1}^{2} M_i\left(\hat{P}_i - \hat{P}\right)^2}{\hat{P}\left(1-\hat{P}\right)}$$

Appropriate adjustment of x_P^2 for clustering effects requires an estimate of the underlying intracluster correlation coefficient ρ, which, under the null hypothesis of no intervention effect, may be assumed to be constant across intervention groups. The required estimate may be obtained by pooling the observations in both groups and then applying the "analysis of variance approach" described by Donner and Klar (1994). Let MSC and MSW denote the pooled mean square errors between and within groups, respectively. Then defining $\overline{m}_{Ai} = \sum_{j=1}^{k_i} m_{ij}^2 \Big/ M_i$, We obtain $\hat{\rho} = \left(MSC - MSW\right)\Big/\left(MSC + \left[m_0 - 1\right]MSW\right)$, where

$$MSC = \sum_{i=1}^{2}\sum_{j=1}^{k} m_{ij}\left(\hat{P}_{ij} - \hat{P}_i\right)^2 \Big/ (k-2)$$

$$MSW = \sum_{i=1}^{2}\sum_{j=1}^{k_i} m_{ij}\hat{P}_{ij}\left(1-\hat{P}_{ij}\right)\Big/(M-k)$$

and $m_0 = \left[M - \sum_{i=1}^{2}\overline{m}_{Ai}\right]\Big/(k-2)$

The value of x_P^2 is then adjusted by applying a correction factor which depends on both $\hat{\rho}$ and the values of the m_{ij}. Letting $C_i = 1 + \left(\overline{m}_{Ai}\right)\hat{\rho}$, the adjusted chi-square statistic with one degree of freedom is given by $x_A^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i\hat{P}\left(1-\hat{P}\right)}$. At $\hat{\rho} = 0, (C_1 = C_2 = 1)$ is it clear that x_A^2 reduces to x_P^2 while if all clusters are of the same size m, it reduces to $x_P^2 / [1+(m-1)\hat{\rho}]$.

This approach may also be used to construct an approximate confidence interval about $(\hat{P}_1 - \hat{P}_2)$. Using the notation above, a two sided 95% confidence interval is given by

$$(\hat{P}_1 - \hat{P}_2) \pm 1.96\sqrt{\frac{C_1 P_1(1-P_1)}{M_1} + \frac{C_2 P_2(1-P_2)}{M_2}}$$

. At $\hat{\rho} = 0, (C_1 = C_2 = 1)$ this expression reduces to the standard confidence interval about a difference between two proportions. However the assumption of a common intracluster correlation coefficient, although guaranteed under the null hypothesis of no intervention effect, may not be appropriate for confidence interval construction. In this case separate estimates of ρ may be used in computing the variance inflation factors C_1 and C_2.

19. Summary

The purpose of this chapter is to provide a basic understanding of methodological issues that must be addressed when investigators decide to randomize intact social units, or clusters of individuals, to different intervention groups. Foremost among these is the justification for randomizing clusters rather than individuals given the loss of statistical efficiency that inevitably arises. The impact of cluster randomization on sample size estimation, the choice of an experimental design, and the approach to the statistical analysis are discussed in detail. Consideration is also given to the unique ethical issues that arise when clusters are selected as the unit of randomization. The chapter closes with suggested guidelines for the reporting of trial results.

20. References

Alexander F., Roberts M.M., Lutz W., Hepburn W. (1989) "Randomization by cluster and the problem of social class bias." *Journal of Epidemiology and Community Health*;43:29-36.

Althabe F., Belizan J.M., Villar J., et al. (2004) “Mandatory second opinion to reduce rates of unnecessary caesarean sections in Latin America: a cluster randomised controlled trial.” *The Lancet* 363:1934-1940.

Altman D.G., Schulz K.F., Moher D., et al. (2001) “The revised CONSORT statement for reporting randomised trials: Explanation and elaboration.” *Annals of Internal Medicine* 134:663-94

Agarwal G.G., Awasthi S., Walter S.D. (2005) “Intra-class correlation estimates for assessment of vitamin A intake in children.” *Journal of Health, Population and Nutrition* 23:66-73.

Atienza A.A., King A.C. (2002) “Community-based health intervention trials: An overview of methodological issues.” *Epidemiologic Reviews;* **24**, 72-79.

Baldwin S.A., Murray D.M., Shadish W.R. (2005) “Empirically supported treatments or Type 1 errors? Problems with the analysis of data from group-administered treatments.” *Journal of Consulting and Clinical Psychology* 73:924-935.

Bass M.J., McWhinney I.R., Donner A. (1986) “Do family physicians need medical assistants to detect and manage hypertension?” *Canadian Medical Association Journal* 134:1247-1255.

Begg C., Cho M., Eastwood S., et al. (1996) “Improving the quality of reporting of randomized controlled trials, The CONSORT statement.” *Journal of the American Medical Association;* **276**, 637-639.

Blair R.C., Higgins J.J. (1986) “Comment on statistical power with group mean as the unit of analysis.” *Journal of Educational Statistics* 11:161-169.

Bland J.M. (2004) “Cluster randomised trials in the medical literature:Two bibliometric surveys.” *BMC Medical Research Methodology*;4: 1-6.

Bellamy S.L., Gibberd R., Hancock L., et al. (2000) “Analysis of dichotomous outcome data for community intervention studies.” *Statistical Methods in Medical Research* 9:135-159.

Campbell M.K., Mollison J., Steen N., et al. (2000) “Analysis of cluster randomized trials in primary care: a practical approach.” *Family Practice* 17:192-196.

Campbell M.K., Elbourne D.R., Altman D.G. for the CONSORT Group. (2004) “CONSORT statement: extension to cluster randomised trials.” *BMJ* 328:702-708.

Campbell M.J., Donner A., Klar N. (2007) “Developments in cluster randomized trials and *Statistics in Medicine*.” *Statistics in Medicine* 26:2-19.

COMMIT Research Group. (1995) “Community intervention trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention.” *American Journal of Public Health* 85:183-192.

Cornfield J. (1978) “Randomization by group: a formal analysis.” *American Journal of Epidemiology* 108:100-102.

Davies M.J., Heller S., Skinner T.C., et al. (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 randomised controlled trial.” *BMJ* 336:491-495.

Divine G.W., Brown J.T., Frazier L.M. (1992) “The unit of analysis error in studies about physicians’ patient care behavior.” *J Gen Internal Medicine* 7:623-629.

Diwan V.K., Wahlström R., Tomson G., Beermann B., Sterky G., Eriksson B. (1995) “Effects of ‘Group Detailing’ on the prescribing of lipid-lowering drugs: A randomized controlled trial in Swedish primary care.” *Journal of Clinical Epidemiology*;**48**, 705-711.

Donner A., Brown K.S., Brasher P. (1990) “A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979 – 1989.” *International Journal of Epidemiology* 19:795-800.

Donner A., Klar N. (1994) “Methods for comparing event rates in intervention studies when the unity of allocation is a cluster.” *American Journal of Epidemiology* 140:279-289.

Donner A., Klar N. (1996) “Statistical considerations in the design and analysis of community intervention trials.” *Journal of Clinical Epidemiology* 49:435-439.

Donner A. (1998) “Some aspects of the design and analysis of cluster randomization trials.” *Applied Statistics* 47:95-114.

Donner A., Klar N. (2000) *Design and analysis of cluster randomization trials in health research*. New York: Oxford University Press.

Donner A., Klar N. (2004) “Pitfalls of and controversies in cluster randomization trials.” *American Journal of Public Health* 94:416-421.

Edwards S.J.L., Braunholtz D.A., Lilford R.J., Stevens A.J. (1999) “Ethical issues in the design and conduct of cluster randomised controlled trials.” *BMJ* 318:1407-1409.

Eldridge S.M., Ashby D., Kerry S. (2006) “Sample size randomized trials: effect of coefficient of variation of cluster size and analysis method.” *Int J Epidemiology* 35:1292-1300.

Eldridge S.M., Ashby D., Bennett C., Wakelin M., Feder G. (2008) “Internal and external validity of cluster randomised trials: systematic review of recent trials.” *BMJ* 336:876-880.

Farquhar J.W., Maccoby N., Wood P.D., et al. (1977) “Community education for cardiovascular health.” *Lancet* 1:1192-1195.

Farr B.M., Hendley J.O., Kaiser D.L., Gwaltney J.M. (1998) “Two randomized controlled trials of virucidal nasal tissues in the prevention of natural upper respiratory infection.” *American Journal of Epidemiology* 128:1162-1172.

Farrin A., Russell I., Torgerson D., Underwood M. on behalf of the UK BEAM Trial Team. (2005) “Differential recruitment in a cluster randomized trial in primary care: the experience of the UK Back pain, Exercise, Active management and Manipulation (UK BEAM) feasibility study.” *Clinical Trials* 2:119-124.

Feldman H.A., McKinlay S.M. (1994) “Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model.” *Statistics in Medicine*; **13**, 61-78.

Feng Z., McLerran D., Grizzle J. (1996) “A comparison of statistical methods for clustered data analysis with Gaussian error.” *Statistics in Medicine* 15:1793-1806.

Fontanet A.L., Saba J., Chandelying V., et al. (1998) “Protection against sexually transmitted diseases by granting sex workers in Thailand the choice of using the male or female condom: results from a randomized controlled trial.” *AIDS* 12:1851-1859.

Grosskurth H., Mosha F., Todd J., et al. (1995) “Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomized controlled trial.” *Lancet* 346:530-536.

Guittet L., Ravaud P., Giraudeau B. (2006) “Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes.” *BMC Medical Research Methodology* 6:17.

Gulliford M.C., Adams G., Ukoumunne O.C., et al. (2005) “Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data.” *Journal of Clinical Epidemiology* 58:246-251.

Haggerty P.A., Muladi K., Kirkwood B.R., Ashworth A., Manunebo, M. (1994) “Community-based hygiene education to reduce diarrhoeal disease in rural Zaire: Impact of the intervention on diarrhoeal morbidity.” *International Journal of Epidemiology*; 23,1050-1059.

Hickman M., McDonald T., Judd A., et al. (2008) “Increasing the uptake of hepatitis C virus testing among injecting drug users in specialist drug treatment and prison settings by using dried blood spots for diagnostic testing: A cluster randomized controlled trial.” *Journal of Viral Hepatitis* 15:250-254.

Hunt M.K., Fagan P., Lederman R., et al. (2008) “Feasibility of implementing intervention methods in an adolescent worksite tobacco control study.” *Tobacco Control* 12(Suppl IV):iv40-iv45.

Jacobs D.R., Luepker R.V., Mittel M., et al. (1986) “Community-wide prevention strategies:evaluation design of the Minnesota Heart Health Program.” *J Chronic Disease* 39:775-88.

Jago R., Baranowski T., Baranowski J.C., et al. (2006) “Fit for life boy scout badge: Outcome evaluation of a troop and Internet intervention.” *Preventive Medicine* 42:181-187.

Johnsson K.O., Berglund M. (2003) “Education of key personnel in student pubs leads to a decrease in alcohol consumption among the patrons: a randomized controlled trial.” *Addiction* 98:627-633.

Kidane G., Morrow R.H. (2000). “Teaching mothers to provide home treatment of malaria in Tigray, Ethiopia: A randomised trial.” *Lancet*; 356, 550-555.

Klar N., Donner A. (1997) “The merits of matching in community intervention trials.” *Statistics in Medicine* 16:1753-1764.

Klar N., Donner A. (2007a) “Ethical challenges posed by cluster randomization.” In: Massaro J, DAgostino RB, Sullivan LM, eds. *Wiley Encyclopedia of Clinical Trials*. United States: John Wiley & Sons, Inc.:1-5.

Klar N., Donner A. “Cluster Randomization.” In: Massaro J, DAgostino RB, Sullivan LM, eds. (2007b) *Wiley Encyclopedia of Clinical Trials.* United States: John Wiley & Sons, Inc.:329-345.

Koepsell T.D., Martin D.C., Diehr P.H., et al. (1991) “Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs; a mixed-model analysis of variance approach.” *Journal of Clinical Epidemiology*;**44**, 701-713.

Kramer M.S., Chalmers B., Hodnett E.D., et al. (2001) “Promotion of breastfeeding intervention trial (PROBIT): A randomization trial in the Republic of Belarus.” *JAMA* 285:413-420.

Kreft I.G.G. (1998) “An illustration of item homogeneity scaling and multilevel analysis techniques in the evaluation of drug prevention programs.” *Evaluation Review* 22:46-77.

Lake S., Kammann E., Klar N., Betensky R. (2002) “Sample size re-estimation in cluster randomized trials.” *Statistics in Medicine* 21:1337-1350.

LaPrelle J., Bauman K.E., Koch G.C. (1992) “High intercommunity variation in adolescent cigarette smoking in a 10-community field experiment.” *Evaluation Review* 16:115-130.

Lasater T.M., Becker D.M., Hill M.N., Gans K.M. (1997) “Synthesis of findings and issues from religious-based cardiovascular disease prevention trials.” *Annals of Epidemiology* 7:S46-S53.

Liang K-Y., Zeger S.L. (1986) “Longitudinal data analysis using generalized linear models.” *Biometrika* 76:13-22.

Mantel N., Haenszel W. (1959) “Statistical aspects of the analysis of data from retrospective studies of disease.” *Journal of the National Cancer Institute* 22:719-748.

Martin D.C., Diehr P., Perrin E.B., Koepsell T.D. (1993) “The effect of matching on the power of randomized community intervention studies.” *Statistics in Medicine* 12:329-338.

Mason S., Knowles E., Colwell B., et al. (2007) “Effectiveness of paramedic practitioners in attending 999 calls from elderly people in the community: cluster randomised controlled trial.” *BMJ* 3335:919.

Meinert C.L. “Long-term drug prevention trials.” (2008) *Clinical Trials* 2:168-176.

Mollison J., Simpson J.A., Campbell M.K., Grimshaw J.M. (2000) “Comparison of analytical methods for cluster randomised trials: an example from a primary care setting.” *J Epidemiology and Biostatistics* 5:339-348.

Moher D., Schulz K.F., Altman D.G. for the CONSORT Group. (2001) “The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomised trials.” *Lancet*; **357**, 1191-1194.

Morgenstern H. “Ecologic studies.” In: Rothman KJ, Greenland S, eds. (1998) *Modern Epidemiology, 2nd Edition*. Philadelphia, PA: Lippincott-Raven:Ch.22.

Murray D.M. (2001) “Statistical models appropriate for designs often used in group-randomized trials.” *Statistics in Medicine* 20:1373-1385.

Murray D.M. (1997) “Design and analysis of group-randomized trials: a review of recent developments.” *Annals of Epidemiology* 7(Supplement):S69-S77.

Murray D.M. (1998) *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Murray D.M, Clark MH, Wagenaar AC. (2000) “Intraclass correlations from a community-based alcohol prevention study: The effect of repeat observations on the same communities.” *Journal of Studies on Alcohol* 61:881-890.

Murray D.M, Pals SL, Blitstein JL, Alfano CM, and Lehman J. (2008) “Design and analysis of group-randomized trials in cancer: A review of current practices.” *J Nat Cancer Inst* 100:483-491.

Nixon R.M., Thompson S.G. (2003) “Baseline adjustments for binary data in repeated cross-sectional cluster randomized trials.” *Statistics in Medicine* 22:2673-2692

O’Brien P.C.,Fleming T.R. (1979) “A multiple testing procedure for clinical trials.” *Biometrics* 35:549-556.

Pan W., Wall M.M. (2002) “Small sample adjustments in using the sandwich variance estimator in generalizing estimating equations.” *Statistics in Medicine* 21:1429-1441.

Parker D.R., Evangelou E., Eaton C.B. (2005) “Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART).” *Contemporary Clinical Trials* 26:260-267.

Peterson A.V. Jr., Kealey K.A., Mann S.L., Marek P.M., Sarason I.G. (2002) “Hutchinson smoking Prevention Project: Long-term randomized trial in school-based tobacco use prevention-results on smoking.” *Journal of the National Cancer Institute*; 92,1979-1991.

Ray W.A., Taylor J.A., Meador K.G., Thapa P.B., Brown A.K., Kajihara H.K., Davis C., Gideon P., Griffin M.R. (1997) “A randomized trial of a consultation service to reduce falls in nursing homes.” *Journal of the American Medical Association*; 278, 557-562.

Roberts C., Roberts S.A. (2005) “Design and analysis of clinical trials with clustering effects due to treatment.” *Clinical Trials* 2:152-162.

Senn S. (1994) “Testing for baseline balance in clinical trials.” *Statistics in Medicine* 13:1715-1726.

Simpson J.M., Klar N., Donner A. (1995) “Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993.” *American Journal of Public Healt*h 85:1378-1382.

Skinner C.S., Arfken C.L., Waterman B. (2000) “Outcomes of the Learn,Share and Live breast cancer education program for older urban women.” *American Journal of Public Health* 90:1229-1234.

Smith P.J., Moffatt M.E.K., Gelskey S.C., Hudson S., Kaita K. (1997) “Are community health interventions evaluated appropriately? A review of six journals.” *Journal of Clinical Epidemiology* 50:137-146.

Sommer A., Tarwotjo I., Djunaedi E., West K.P. Jr, Loeden A.A., Tilden M.L. and the ACEH Study Group. (1986) “Impact of vitamin A supplementation on childhood mortality.” *Lancet* 1:1169-1173.

Stanton B.F., Clemens J.D. (1987) “An educational intervention for altering water-sanitation behaviors to reduce childhood diarrhea in urban Bangladesh.” *American Journal of Epidemiology*;125, 292-301.

Thompson S.G., Pyke S.D.M., Hardy R.J. (1997) “The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques.” *Statistics in Medicine*;16, 2063-2980.

Torgerson D.J. (2001) “Contamination in trials: is cluster randomisation the answer?” *BMJ* 322:355-357.

Tuomilehto J., Nissinen A., Salonen J.T., Kottke T.E., Puska P. (1980) “Community programme for control of hypertension in North Karelia, Finland.” *Lancet* 2:900-903.

Turpeinen O., Karvonen M.J., Pekkarinen M., Miettinen M., Elosuo R., Paavilainen E. (1979) “Dietary prevention of coronary heart disease: the Finnish mental hospital study.” *International Journal of Epidemiology* 8:99-118.

Ukoumunne O.C. and Thompson S.G. (2001) “Analysis of cluster randomized trials with repeated cross-sectional binary measurements.” *Statistics in Medicine* 20:417-433.

Varnell S.P., Murray D.M., Janega J.B., Blitstein J.L. (2004) “Design and analysis of group-randomized trials: A review of recent practices.” *American Journal of Public Health* 94:393-399.

Walsh M.W., Hilton J.F., Masouredis C.M., et al. (1999) “Smokeless tobacco cessation intervention for college athletes:results after one year.” *American Journal of Public Health* 89:228-234.

West K.P., Pokhrel R.P., Katz J., et al. (1991) “Efficacy of vitamin A in reducing preschool child mortality in Nepal.” *Lancet* 338:67-71.

White I.R., Thomas J. (2005) “Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis.” *Society for Clinical Trials* 2:141-151.

Whiting-O’Keefe Q.E., Henke C., Simborg D.W. (1984) “Choosing the correct unit of analysis in medical care experiments.” *Medical Care* 22:1101-1114.

Zou G., Donner A., Klar N. 2005) “Group sequential methods for cluster randomization trials with binary outcomes.” *Clinical Trials* 2:479-487.

21. Author Biography

Allan Donner, MSc, PhD is Professor in the Department of Epidemiology and Biostatistics at the University of Western Ontario. His methodological research focuses on the design and analysis of clinical studies, with a special interest in cluster randomization trials. He is a co-author of the text "Design and Analysis of Cluster Randomization Trials in Health Research". Dr. Donner has served on the Steering Committee of multi-national trials sponsored by the World Health Organization and the European Commission, is a consultant to the International Vaccine Institute based in Korea, and is a member of the Expert Advisory Committee on Bioavailability and Bioequivalence at Health Canada. He has presented invited talks and workshops for the Society for Clinical Trials, the Drug Information Association, and the Biometric Society.