ANALYSIS OF CLUSTER RANDOMISED TRIALS WITH DISCRETE OUTCOMES

Maupi Eric Letsoalo

BSc. (Honours)

Submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

Mathematical Statistics

in the

Department of Statistics

Faculty of Natural and Agricultural Sciences

University of Pretoria

2004

Declaration

I declare that **Analysis of Cluster Randomised Trials with Discrete Outcomes** is my own work and that all the resources used or quoted in the study have been indicated and reflected by means of complete references.

…………………………………………….

Maupi Eric Letsoalo

For my daughter, **Mampe,** and my son, **Maupe**

Acknowledgements

My sincere thanks and appreciation are due to all those people who contributed in so many ways to facilitate the completion of this thesis.

I thank

- Medical Research Council of South Africa (MRC), for financial support.

- Prof P J Becker (MRC), for his motivating ‘*sjambok*’, and for his professional and intellectual guidance.

- Gender and Health Unit (MRC), for allowing me to use their data set.

- My wife Mamosana for consistently ‘*smashing-my-head*’ and for her famous reminder *“Know you don’t have much time”*, her moral support, and for giving me ‘*some reasons*’ to study.

- My colleagues at MRC, especially S. D. Moti and E. Gomes (Bio-Statistics Unit) and M. Mefane (Information Service Division), for their technical support.

- Dr K M Chuene (University of the North, UNIFY Section) for her professional inputs and ‘*sisterly guidance*’, and my brother M. D. Letsoalo (University of South Africa (UNISA), English Department), for encouragement and moral support.

I would like to express my sincere gratitude to:

- ► My supervisor, Prof H T Groeneveld, for his professional guidance, and,

- ► My co-supervisor, Dr J B Levin, for his constant intellectual support, encouragement, patience and professional guidance.

Abstract

A cluster randomisation trial (CRT) is an experiment in which a researcher randomises intact clusters (e.g. families, schools, communities) rather than individuals in experimental trials. Reasons for adopting cluster randomisation trials are diverse. In this thesis, issues such as the choice of design, sample size calculations and approaches to the analyses, that arise in conducting such trials are discussed, with the emphasis on trials with binary outcomes.

Three approaches that are widely employed in CRT's, namely the aggregated approach, the disaggregated approach and the adjusted chi-square approach, were employed on our data sets and it was found that in some cases the methods gave differing results. In particular, the disaggregated analysis is able to adjust for individual level covariates. Therefore, gaining power, and sometimes detect statistically significant differences where the other methods have not been able to.

.

Table of Contents

List of Figures

List of Tables

APPENDIX

Chapter 1

Introduction

1.1 Cluster Randomised Trials

A cluster randomisation trial (CRT), according to Donner and Klar (2000) is a trial or experiment in which clusters or intact social units of individuals, rather than independent individuals, are randomised to different intervention groups. A CRT is also known as a group randomised trial, group allocation, ‘community randomised’ trial or ‘community intervention’. According to Koepsell (1998) quoted by Donner and Klar (2000), “this may be, in part, because many of the methodological issues involved in community trials stem from the group-allocation feature and do not depend on group size”. Longford (1993) states “Clustering is an omnipresent feature in our everyday lives”. Akoumunne and Thompson (2001) state “Cluster randomised trials are implemented when theoretical, ethical or practical considerations preclude the use of traditional clinical trials in which individual subjects are randomised”.

Cluster randomisation trials are becoming increasingly important in health technology assessment. That is, cluster randomised trials have come to be an important tool in the evaluation of non-therapeutic interventions, including lifestyle modifications, educational programmes and innovations in the provision of health care. In such trials, the units of randomisation are diverse. They range from relatively small clusters such as families or households, to entire neighbourhoods or communities, but also include work places, hospital wards, classrooms and medical practices.

An important aspect of cluster (community) randomisation is that inferences are often made at the individual level while the unit of randomisation is the group.

1.2 Structure of the dissertation

This dissertation is divided into five chapters, with the main parts embodied in chapters 2 to 4. Chapter 2 gives a detailed explanation of cluster randomisation trials, and explicitly introduces the Completely Randomised, Stratified and Matched designs as the three main study designs for cluster randomised trials. The two types of models (aggregated, disaggregated) are also introduced, as models employed in the analysis of continuous outcomes. The possible methods of analysis of binary (discrete) outcomes are discussed in Chapters 3.

Chapter 4 presents the application of aggregated analysis and disaggregated analysis, to data sets, respectively, with discussions and recommendations given in Chapter 5.

1.3 Description of data sets

The data sets that have been used throughout this dissertation are:

a. Randomised cluster trial of "Stepping Stones" behavioural intervention for HIV

"Stepping Stones" is a workshop series of 14 sessions, which cover HIV, sexual, and reproductive health, gender, communication and relationship skills. It was developed in Uganda and has been adapted and used in many developing countries. The South African adaptation was made in 1998 and has since been used widely in the country.

The overall objective of the trial is to determine the effectiveness of "Stepping Stones" in reducing the transmission of HIV and promoting HIV risk reducing practices among poor rural adults in South Africa. The study seeks to determine the effectiveness of "Stepping Stones" in changing aspects of gender dynamics in relationships, including violent practices, and measure the impact of this on HIV risk reduction. It also aims to determine the magnitude of risk of HIV attributable to sexual practices including condom use with main partner, condom use with casual partners, having multiple partners and strong male dominance of relationships.

The study design is a cluster randomised trial set in the former Transkei in the Eastern cape (regions D and E).

Each cluster consists of a pair of groups, one male and one female, in a single geographical area. These two groups each have 20 participants, which is the maximum number who can receive “Stepping Stones” together.

A total of 70 clusters have been randomised to one of the two arms: In the intervention arm (35 clusters), “Stepping Stones” will be provided to each group (male and female) by a facilitator and co-facilitator. The group will work through 14 content-based sessions of two to three hours each, meeting twice a week over eight weeks. The control arm (35 clusters) will receive a single two-hour session on HIV, STDs and safer sexual practices, and will be given “Stepping Stones” at the end of study. The clusters were stratified before randomisation, with one township stratum and seven rural strata, with stratification based on geographical location. The stratum sizes range from 6 to 14 clusters per stratum.

The study design is a cohort study within the ramdomised trials, with the same individuals being followed up throughout the study. This is applicable due to the nature of the intervention, as well as the objective of the study related to HIV incidence.

The cluster sizes are constant i.e., each cluster has 20 men and 20 women.

b. ‘**DOTSPack’: the impact of a novel drug-packaging concept on tuberculosis treatment delivery, patient adherence and treatment outcome.**

The overall aim of this cluster-randomised trial was to evaluate the impact on treatment delivery of pre-packed, blister calendar treatment cards contained in patient-specific treatment packs. These ‘DOTSPacks’ were colour-coded according to treatment weight categories and contained all necessary patient information for proper management. The main objective of the study was to evaluate the impact of DOTSPacks on patient adherence to treatment and consequently on cure and treatment interruption rates.

The study population consisted of confirmed new adult tuberculosis patients registered sequentially at TB treatment units in the seven health districts of the Highveld region of the Mpumalanga Province. There were 42 treatment units included in the study.

The study was a cluster randomised controlled trial with the clusters being treatment units, and with randomisation stratified by health district, to one of the two arms namely an intervention arm (where treatment was delivered by the DOTSPacks) and a control arm (where treatment was delivered in the conventional way). All 22-treatment units were randomised to the DOTSPack arm, and 20 to the control arm. This resulted in a total of 284 subjects in the intervention arm and 220 subjects in the control arm.

The design variables in this study are district, clinic (cluster) and group (DOTSPack or not). The outcome variable is binary and looks at favorable outcomes (cured or treatment completed) versus unfavorable outcome (unknown, treatment interruption or treatment failure).

In addition, the following covariates were recorded: age (in years), gender, whether the decision to treat was based on bacteriology, baseline sputum smear and whether or not the patient had extra-pulmonary TB.

Chapter 2

Literature review

2.1 Introduction: Cluster Randomised Trials

Clinical trials typically involve the randomisation of individual subjects to the intervention or group control, whereas cluster randomised trials are characterised by the randomisation of groups or clusters of individuals to treatment groups (Bennet *et al.* 2002).

Cluster Randomised Trials have been widely used for evaluation of health care and education strategies, where medical practices, schools, or even entire communities are selected as the unit of randomisation (Donner, 1997). Outcomes are measured on individuals within those clusters. A distinguishing feature of cluster randomised trials is that the response of individuals within a cluster may be correlated, which results in the statistical efficiency of cluster randomised designs being lower than that of individually randomised designs (Bennett *et al.*, 2002). Cornfield (1978) also states that such allocation schemes are less efficient, in a statistical sense, than designs that randomise individuals to intervention groups.

The degree of similarity among responses within a cluster is measured by a parameter known as the intracluster (or intraclass) correlation coefficient ρ. According to Klar and Donner (1997) the intracluster correlation coefficient depends upon both the outcome variable and the type of cluster, and tends to be larger in smaller clusters, although the relationship of ρ to cluster size tends to be non-linear (Donner, 1997; Mickey and Goodwin, 1993; Murray *et al.*, 1994). The intracluster correlation may be interpreted as the standard Pearson correlation coefficient between any two responses in the same cluster.

If we add the (additional) assumption that the intraclass correlation cannot be negative, ρ may also be interpreted as the proportion of overall variation in response that can be accounted for by the between-cluster variation (Donner and Klar, 2000). A positive

intracluster correlation coefficient implies that the variation between observations in different clusters exceeds the variation within clusters; hence it can be claimed that the design is characterised by 'between-cluster variation' (Cochran and Cox, 1957; Donner and Klar 2000; Cochran, 1997). This is also equivalent to stating that clusters cannot be assumed to be 'interchangeable' with regard to the experimental endpoints (Klar and Donner, 2000), and the underlying reasons for lack of interchangeability or variation between clusters will differ from trial to trial, but in practice may include:

i. Subject selection, where individuals are in a position to choose the cluster to which they belong.
ii. The influence of the covariates at the cluster level, where all individuals in a cluster are affected in a similar manner as a result of sharing exposure to a common environment.
iii. The tendency of infectious diseases to spread more rapidly within families than between families or throughout communities.
iv. The effect of personal interactions among cluster members who receive the same intervention.

Without extensive empirical data, it is usually impossible to distinguish among the potential reasons for between-cluster variation (Donner and Klar, 2000). The intracluster correlation results in the statistical efficiency of cluster randomised designs being lower than that of individually randomised designs. That is, between-cluster variation invariably leads to a reduction in the effective sample size for the trial, where the size of the reduction increases with both the magnitude of ρ and the average cluster size. And this leads to a loss of precision in estimating the effect of the trial intervention. Hence Bennett *et al.* (2002) state that a large sample size is necessary for equivalent power. Failure to take account of the intra-cluster correlation in analysis does not bias the point estimate of the intervention effect, but leads to falsely narrow confidence intervals and an inappropriately small *p*-value.

Cluster randomised controlled trials are attractive in settings in which individual randomisation is difficult or impossible (Isaakidid and Loannidid, 2003). Donner (1997) emphasises the fact that the justification for adopting a cluster randomisation design must rest on practical considerations such as the desire to control costs or attempting to minimise experimental contamination, and ethical considerations and should always be stated explicitly. As stated by MRC-UK (2002), there are several key reasons to consider cluster randomised trials:

i. The intervention to be studied is (itself) delivered to and affects groups of people rather than individuals.
ii. The intervention is targeted at health professionals with the aim of studying its impact on patient outcomes.
iii. An intervention is given to individuals but might affect others within that cluster, i.e., there will be contamination. The following figure explains the concept of contamination explicitly.

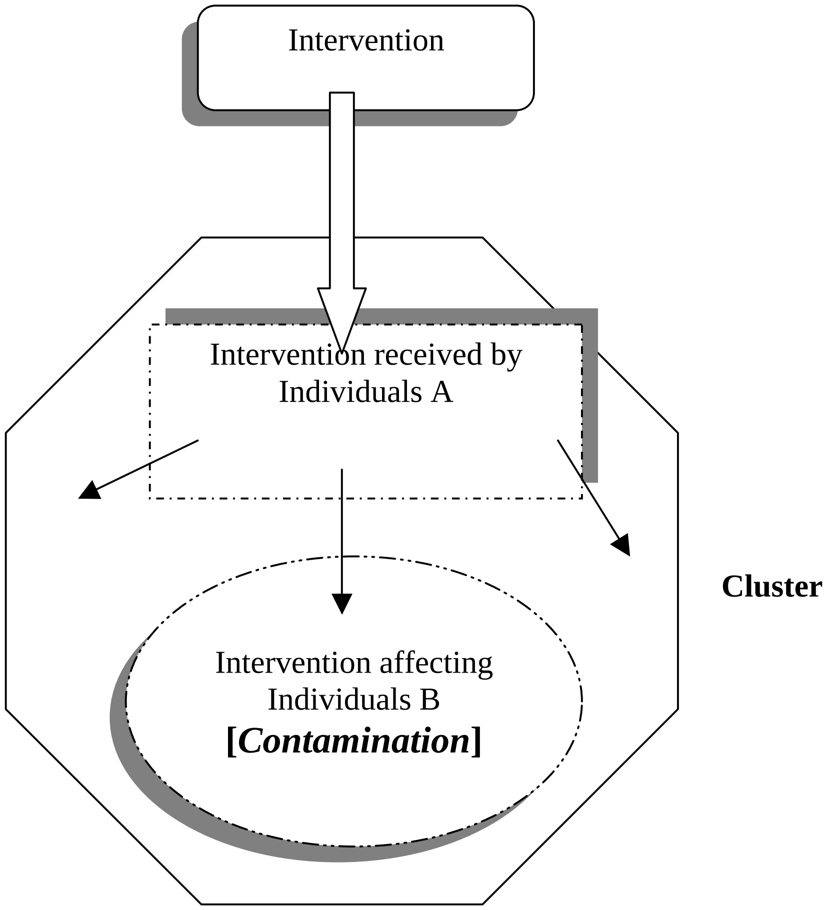

Figure 1: Illustration of 'contamination'

The point of concern is 'the need to minimise or remove contamination'. As Donner and Klar (2000) put it, "the trade-off between the lower precision associated with cluster randomisation and the potential contamination bias associated with individual randomisation is a factor that frequently arises in the decision to adopt a cluster randomisation design". When the contamination is expected only to reduce the strength, effect or value of intervention, then the net effect of this trade-off can be estimated quantitatively.

A simple model that may help in the choice of design that might well account for such attenuation is to assume that a certain portion of control group subjects will, through an extreme form of contamination, *behave* like experimental group subjects. Therefore, the sizes of the clusters randomised, the value of ρ and the expected degree of contamination inform the resulting decision to adopt cluster or individual randomisation.

According to Donner and Klar (2000) the selection of a cluster randomisation design does not guarantee that the threat of contamination will be entirely removed. However this threat can sometimes be further minimised, by implementing the study in a geographic area in which only distinct and well-separated clusters of subjects are recruited.

In most studies, “the investigators choose to adopt a cluster randomisation design in order to avoid ethical, logical or methodological problems that might otherwise have arisen” (Donner and Klar, 2000).

Fisher’s classical theory of experimental design assumes, without exception, that the experimental unit that is randomised is also the unit of analysis. The challenges of cluster randomisation arise because inferences are frequently intended to apply at the individual level, while randomisation is at the cluster level.

As explained earlier, if individual-level analyses are to be performed, the lack of statistical independence among members in a cluster invalidates standard approaches to both the estimation of the sample size and the analysis of the trial data, with standard methods of statistical analysis tending to bias the observed p-value downwards.

It may also be added, as Cornfield (1978) put it, that “randomisation by cluster accompanied by an analysis appropriate to randomisation by individual is an exercise in self-deception, however, and should be discouraged”. On the other hand, randomisation by cluster accompanied by a sample size assessment appropriate to randomisation by individual is also an exercise in self-deception. The former may lead to type I error substantially above that which is planned for, while the latter may lead to a substantially elevated type II error (Donner and Klar, 2000).

2.2 Estimating the intra-cluster correlation coefficient

In quantifying the effect of clustering, we consider a trial in which *k* clusters of sizes *m* are randomly assigned to each of the experimental group and a control group. The following figure explains the situation explicitly.

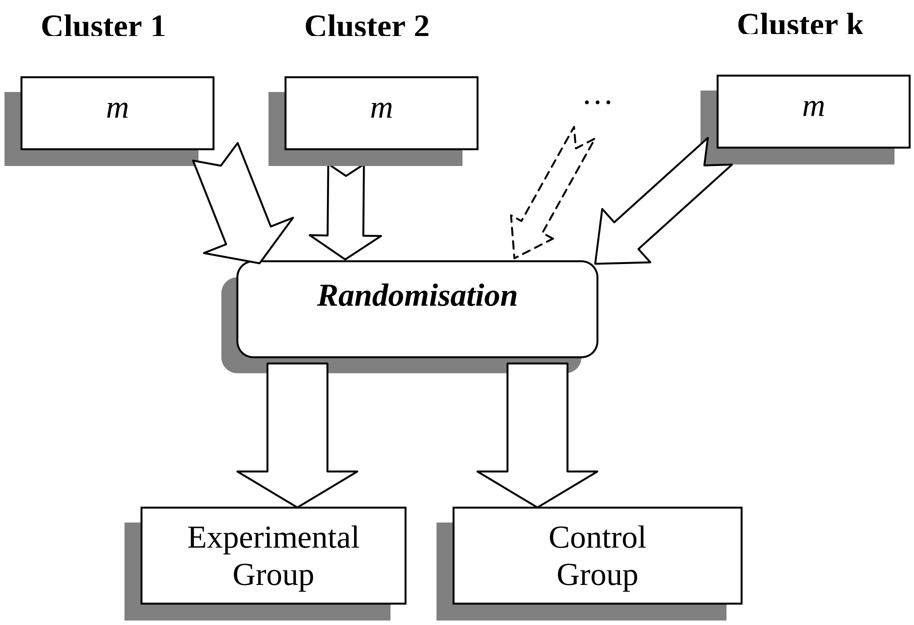

Figure 2: Quantifying the clustering effect

If we measure some variable *Y* on all *km* observations (individuals) such that these measurements are either continuous or discrete and are not necessarily normally distributed, then the error in measurement of members of a cluster are independent, and can be expressed by the within-cluster variance, σ^2 (Wears, 2000). The variance of the jth cluster's mean would then be:

$$Var(\overline{Y}_j) = \frac{\sigma^2}{m}. \tag{1}$$

If the primary aim of the trial is to compare the group means on a normally distributed variable *Y* having a common but unknown variance σ^2, then the estimates of population means μ_1 and μ_2 are sample means $\overline{Y}_1$ and $\overline{Y}_2$ for experimental and control groups, respectively.

Donner and Klar (2000) state a well-known result in cluster sampling (see e.g. Cochran, 1997) that the variance of each of these means is given by

$$Var(\overline{Y}_i) = \frac{\sigma^2}{km}[1+(m-1)\rho], \quad i = 1,2 \qquad \textbf{(2)}$$

where

$\rho = \dfrac{\sigma^2{}_A}{\sigma^2}$ and $\sigma^2{}_A$ are the intracluster correlation coefficient and the variance component among clusters, respectively.

We note that

1. At $\rho = 0$, $Var(\overline{Y}_i) = \dfrac{\sigma^2}{km}$ which is the standard expression for the variance of a sample mean under simple random sampling. This case corresponds to that of statistical independence among members of the cluster.
2. At $\rho = 1$, $Var(\overline{Y}_i) = \dfrac{\sigma^2}{k}$ corresponds to total dependence. In this case all the responses in a cluster are identical, so that the total information supplied by the cluster is no more than that supplied by a single member. In this case the effective sample size, which is generally given by $\dfrac{m}{1+(m-1)\rho} = \dfrac{m}{IF}$, equals 1. The factor $IF = 1+(m-1)\rho$, called the *inflation factor* or *design effect,* is the factor by which the total sample size must be increased if a cluster design is to have the same statistical power as an individual design and can also be used in statistical inference to correct confidence intervals and *p*-values (Cochran, 1977; Donner *et al.*, 1981; Wears, 2000; Donner and Klar, 2000).

In case of unequal cluster sizes, the substitution of the average cluster size $\overline{m}$ for *m* in *IF* will slightly underestimate the variance of $\overline{Y}_i$, unless the variation in cluster sizes is large (Donner and Klar, 2000; Donner *et al*., 1981).

We note, with the assumption that $\rho \geq 0$, that the intraclass correlation represents the ratio of between-cluster variability to total variation as:

$$\rho = \frac{\sigma_A^{\ 2}}{\sigma_A^{\ 2} + \sigma_W^{\ 2}} \tag{3}$$

where,

$$\sigma_A^{\ 2} + \sigma_W^{\ 2} = \sigma^2,$$

$\sigma_W^{\ 2}$ and $\sigma_A^{\ 2}$ are within and between cluster components respectively (Donner and Klar, 2000). The variance of the individual group average can be written as:

$$Var(\bar{Y}_i) = \frac{\left(\frac{\sigma^2}{m}\right)}{k}. \tag{4}$$

The standard one-way analysis of variance (ANOVA) may be used to determine an estimate of ρ. Considering the case where a single sample has *k* clusters, each of size *m*, ANOVA estimate of ρ is given by;

$$\hat{\rho} = \frac{S_A^{\ 2}}{S^2{}_A + S^2{}_W} = \frac{MSC - MSW}{MSC + (m-1)MSW}, \tag{5}$$

where $S_A^{\ 2} = \frac{(MSC - MSW)}{m}$ and $S_W^{\ 2} = MSW$ are sample estimates of $\sigma_A^{\ 2}$ and $\sigma_W^{\ 2}$ respectively, and *MSC* and *MSW* are mean square errors among and within clusters respectively.

For variable cluster sizes $m_j, j = 1,2,\ldots,k$; m in **(2)** is replaced by:

$$m_0 = \left(\frac{1}{k-1}\right)\left(M - \sum_{J=1}^{K}\frac{m_j^{\ 2}}{M}\right), \tag{6}$$

with $M = \sum_{j=1}^{k} m_j$ being the number of individuals in the sample. According to Donner and Koval (1980) $\hat{\rho}$ can be shown to be a more accurate estimator of ρ, particularly when the cluster sizes are highly variable. Otherwise for clusters of the same size *m*, $\hat{\rho}$ is nearly identical to the ‘pairwise’ estimator of ρ determined from the usual Pearson product-moment correlation over all possible pairs of observations that can be generated within clusters.

If, in ***(2)***, σ^2 is substituted by $P(1-P)$, where P is the probability of success then ***(2)*** provides an expression for the variance of a sample proportion under clustering. To be precise, if Y is the binary response in ***(2)*** then the estimator of ρ may be written as:

$$\hat{\rho} = 1 - \frac{\sum_{j=1}^{k} m_j \hat{P}_j \left(1 - \hat{P}_j\right)}{k\left(\overline{m} - 1\right)\hat{P}(1 - \hat{P})} \qquad (7)$$

where

$\hat{P}_j$ is the proportion of success in the j^{th} cluster, $j = 1,2,\ldots,k$, and

$\hat{P} = \frac{\sum_{j=1}^{k} m_j \hat{P}_j}{\sum_{j=1}^{k} m_j}$ is the overall proportion of successes; and

$\overline{m} = \frac{M}{k}$ is the average cluster size.

It is worth noting that $\hat{\rho}$ is a version of the kappa statistic used to estimate the degree of inter-observer agreement among raters with respect to a binary trait measured repeatedly on each of k subjects (Donner and Klar, 2000). Shoukri (2004) gives a more detailed explanation and demonstration of measures of agreement between observers.

For completely independent observations, the within-cluster variance equals the ordinary variance.

2.3 The most used experimental designs

There are two main approaches to randomisation in cluster randomisation trials, which are unrestricted allocation and restricted allocation. Under unrestricted allocation we have completely randomised allocation, in which clusters from a single pool, with no pre-stratification or matching according to baseline characteristics, are allocated to treatment groups. Treatments are the set of circumstances created for the experiment in response to the research hypotheses, and they are the focus of the investigation (Kuehl, 2000). Under restricted allocation, clusters are first divided into strata according to baseline characteristics and then allocated to groups within the designated strata (MRC-UK, 2002). The diagram that follows attempts to demonstrate/represent the breakdown of cluster randomisation trial.

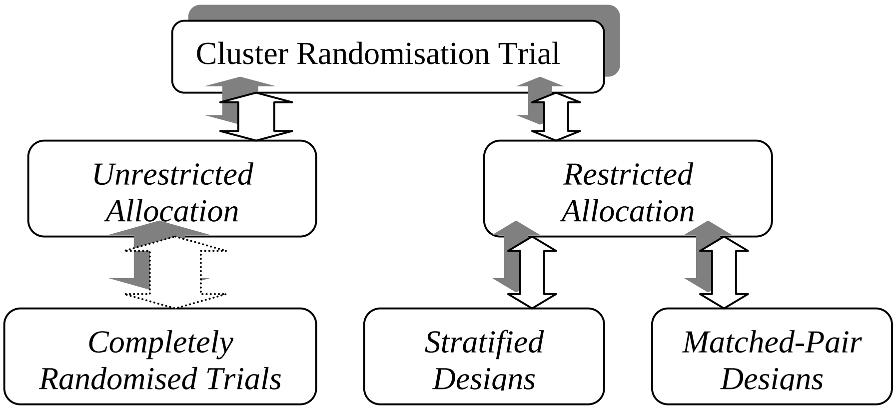

Figure 3: The breakdown of CRTs

The aim of restricted allocation (stratification or matching) is to provide groups that are more evenly balanced with respect to baseline characteristics. Statistical power will also be increased provided the baseline characteristics selected as stratifying factors are strong predictors of outcome (MRC-UK, 2002). These baseline variables may be related to the individuals within clusters or to the characteristics of clusters.

There are two main types of allocation/design under restricted allocation, namely stratified and matched-pair designs. Stratified allocation is when there are more than two clusters in a stratum; within each stratum clusters are allocated to treatment groups by simple or block randomisation.

Simple randomisation is a procedure that uses a code to assign all patients to receive one of the two (or more) different medicine treatments at the start of the trial, i.e. prior to treatment (Spilker, 1991). This code is usually generated prior to the initiation of the trial and usually assigns treatments based on the order of patients, units or clusters admitted to the trial.

Blocking is a means of reducing and controlling experimental error variance in order to achieve more precision (Kuehl, 2000), for the objective is to have precise comparisons among treatments in research studies. In the block randomisation method, a block size is chosen, and the number of patients or number of clusters assigned to each treatment is proportional within the block size chosen.

The main advantage of using block randomisation is that if the trial does not enroll the full number of clusters/individuals expected, there will still be an equal or approximately equal number of patients in each treatment group for a balanced block and proportional number of patients in each group for a proportional block randomisation (Spilker, 1991).

Blocking stratifies experimental units into homogeneous groups, or similar units. If the blocks are small then the subjects within each block will be more homogeneous, and according to Kuehl (2000: 264) general categories of successful blocking criteria are:

i. Proximity (neighboring – in agronomic field experiments)
ii. Physical characteristics (age and/or weight)
iii. Time, and
iv. Management of tasks in the experiment

The blocking variable is easily incorporated into an experimental design, simply by treating it as another factor. However, in general, no interaction between the levels for a blocking variable and the levels of a genuine factor is sensible (Cochran and Cox, 1957).

According to MRC-UK (2002) stratification by cluster size ensures balanced-sized intervention groups when cluster sizes are uneven. The matched-pair design is a special case of a stratified design in which one of the two clusters in each stratum is randomly assigned to each intervention.

Thus there are three designs, which have been frequently used in cluster randomisation trials. A recent review (Donner *et al*., 1990) reveals that investigators tend to adopt one of these principal designs, which are described in more detail below.

2.3.1 Completely randomised trials

Completely randomised trials involve no pre-stratification or matching of clusters according to baseline characteristics and are most suited to trials randomising a fairly large number of clusters (Donner and Klar, 1994). Randomised trials are accepted as the golden standard for the evaluation of new health intervention (Hayes and Bennett, 1999).

Completely randomised trials ensure that each cluster has an equal probability of being assigned to each of the intervention groups. That is, all experimental units (in this case clusters) are allocated at random among all of the treatments (or specific experimental condition applied to the units). The following figure explains the concept/principle of a completely randomisation experiment.

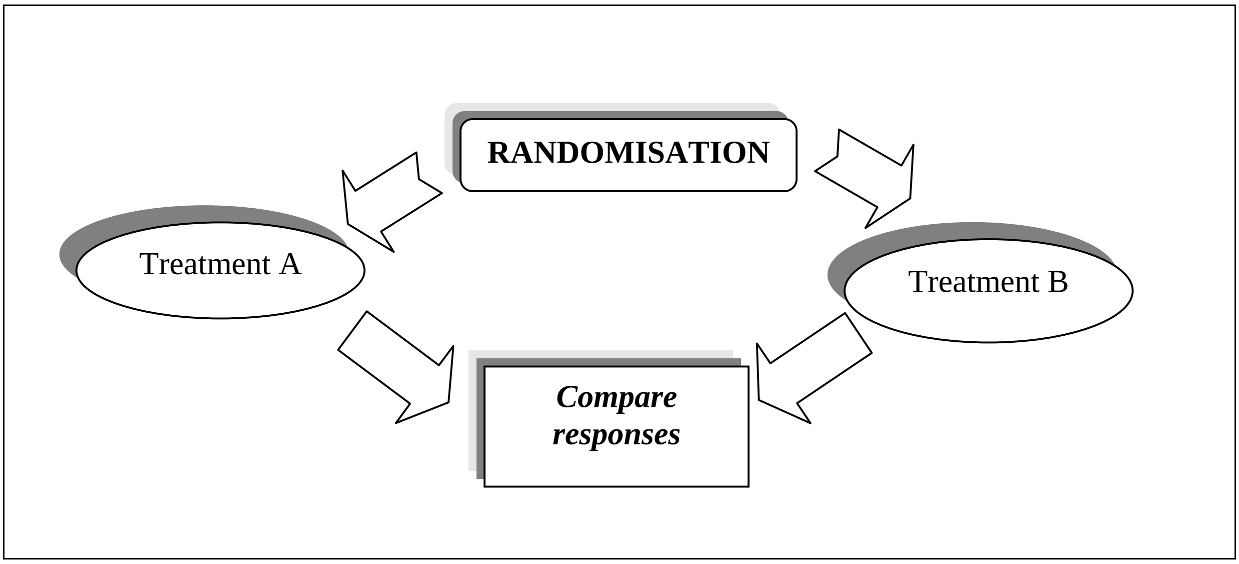

Figure 4: Randomised comparative experiment

Kirk (1995) describes an experimental unit as "that entity that is assigned to an experimental condition independently of such entities." Kuehl (2000) gives a similar definition of experimental unit as "the physical entity or subject exposed to the treatment independently of other units". Randomisation is the best way of ensuring that both known and unknown factors that may independently affect the outcome of an intervention are likely to be distributed evenly between the trial groups.

If a completely randomised trial, as an approach, is used for few clusters, the treatment groups are likely to be unbalanced with respect to baseline characteristics. In this case some matching or stratification in the design is usually advisable, '*to reduce the level of between-cluster variation and hence improve precision*' (MRC-UK, 2002). This is because of the well-accepted design principle that stratification is most effective in small studies, a principle that is particularly relevant to cluster randomisation trials, where the effective sample size may be much less than the actual number of individuals enrolled (Donner and Klar, 2000; Pocock, 1983).

2.3.2 Matched pair designs

Matched pair designs are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control. Matched designs are frequently used in community intervention studies (Martin *et al*., 1993). If there are many known risk factors requiring a finer stratification, a pair-matched design can be performed by matching similar clusters and randomly assigning one member of each match to each treatment (Donner and Klar, 1994).

The main advantage of this design is its potential to provide very tight and explicit balancing of potentially important prognostic factors at baseline (Donner and Klar, 2000); this may enhance the perceived credibility of the trial conclusions.

The two primary reasons for matching are to increase the power of the test or the precision of estimates, and to reduce bias in non-randomised studies (Martin *et al.*, 1993). A matched design gains power when the matched communities are more alike than would be expected due to chance.

The ability of the matched-pair design to create intervention groups that are perceived to be comparable at baseline requires a consensus as to which potential matching factors are the most important. Unless each cluster in a matched pair appear to be somewhat similar, it is unlikely that the results of the research will be accepted as valid.

There are some analytic limitations of these designs that have received attention (Martin *et al.*, 1993; Klar and Donner, 1997). According to MRC-UK (2002) and Donner and Klar (2000) these approaches should be used cautiously because of several important limitations. Thus:

i. It may be difficult to find matching variables to create distinct pairs;

ii. Although matching may reduce the variance of the intervention effect, the resulting increase in power be reduced or even cancelled out by a loss of information on between-cluster variability, especially with small numbers of matched pairs;

iii. It is difficult to estimate the intra-cluster correlation coefficient from data arising from a matched-pair design without making special assumptions; and

iv. If a particular cluster drops out of the study the entire stratum will be eliminated in the analysis, so power is reduced.

2.3.3 The stratified design

The stratified design is an extension of the matched-pair design in which several clusters, rather than just one, are randomly assigned within strata to each of the intervention and control groups (Donner and Klar, 2000). As Hsieh (1998) puts it, "*A stratified design consists of clusters pre-stratified with respect to important baseline characteristics and then, within strata, randomly assigned to either the intervention or the control group*". Consequently, the between cluster variation can be estimated directly since the cluster effect can be separated from both the intervention effect and the stratum effect (MRC-UK, 2002).

Stratification purports to decrease baseline differences between the two groups, intervention and control groups (Hsieh, 1998). The choice of stratification variables may be dictated as much by practical as by scientific considerations (Donner and Klar, 2000). Since precision is increased when there are equal numbers of subjects in treatment groups, stratification by cluster size is recommended whenever the number of subjects per cluster is highly variable (Donner and Klar, 1994).

2.4 Sample size determination

In cluster randomisation trials the responses of individuals within clusters tend to be correlated. Therefore, the information provided by a given sample size in a cluster-randomised trial is generally less than in an individually randomised trial. That is, the statistical power of cluster-randomised trials may be substantially less than that of a similar-sized individually randomised trial, and this has to be taken into account in determining the sample size in cluster randomisation trials (Hayes and Bennett, 1999). In the design stage of cluster randomisation trials, sample size and power determination/calculations performed by using standard methods which assume independence among responses of cluster members lead to underpowered studies (Lake *et al*., 2002).

Many studies and/or reports have discussed sample size calculations, and "focused on a particular type of outcome variable (binary, continuous or person-year rates), relate to either matched or unmatched designs" (Hayes and Bennett, 1998).

2.4.1 Completely Randomised Trials

i. Comparison of means

We consider a situation in which *k* clusters of *m* subjects are randomly allocated to either control or experimental groups (see Figure 2). Suppose Y, the primary response variable for an individual, is normally distributed with common, but unknown variance,

$$\sigma^2 = \sigma_A{}^2 + \sigma_W{}^2, \tag{8}$$

where $\sigma_A{}^2$ and $\sigma_W{}^2$ denote the variance among and within clusters, respectively.

If the null hypothesis $H_0 : \mu_1 = \mu_2$ is to be tested at the two-sided 100α percent level of significance with power $1-\beta$, according to Donner *et al*.(1981), the number of subjects required per intervention group is given by:

$$n = \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \left(2\sigma^2\right)\left[1+(m-1)\rho\right]}{\left(\mu_1 - \mu_2\right)^2}. \tag{9}$$

Equivalently, the number of clusters required per group is:

$$k = \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \left(2\sigma^2\right)\left[1 + (m-1)\rho\right]}{m\left(\mu_1 - \mu_2\right)^2}, \tag{10}$$

where,

$\rho = \dfrac{\sigma_A{}^2}{\sigma_A{}^2 + \sigma_W{}^2}$ is the intracluster correlation,

$Z_{\alpha/2}$ is the two sided critical value of the standard normal distribution corresponding to the error rate α, and Z_β is the critical value corresponding to β, with $\mu_1 - \mu_2$ denoting the size of the difference to be detected, and μ_1 and μ_2 being the true (population) means.

If $\rho = 0$, the case corresponding to that of statistical independence among members of the cluster, then the number of subjects required per intervention group becomes:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2\left(2\sigma^2\right)}{\left(\mu_1 - \mu_2\right)^2}. \tag{11}$$

We note that for one-sided tests $Z_{\alpha/2}$ is substituted by Z_α. The use of critical values $Z_{\alpha/2}$, Z_β in formulae/equation (**11**) rather than critical values $t_{\alpha/2}$, t_β corresponding to the t-distribution will underestimate the required sample size unless the degrees of freedom are large (Donner and Klar, 2000). In most community intervention studies, the number of clusters is small and the population standard deviation is estimated by sample standard deviation S_d. The t-test is more appropriate in testing the hypothesis of no intervention effect.

Given the same situation as explained above, except that $\sigma_1{}^2$ and $\sigma_2{}^2$ are variances of the response variable in the experimental and control groups, respectively, i.e. if the variances $\sigma_1{}^2$ and $\sigma_2{}^2$ within groups are not equal, then according to Hayes and Bennett (1999), the number of subjects per intervention group is:

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \left(\sigma^2{}_1 + \sigma_2{}^2\right)\left[1 + (m-1)\rho\right]}{\left(\mu_1 - \mu_2\right)^2}, \tag{12}$$

or for $\rho = 0$;

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 \left(\sigma^2{}_1 + \sigma_2{}^2\right)}{\left(\mu_1 - \mu_2\right)^2}, \tag{13}$$

In case of unequal cluster sizes, according to Donner *et al*. (1981), m is replaced by an estimate of the average cluster size $\overline{m} = M / k$, where M is the total number of individuals in the sample, in **(09***)***.** This approach will tend to slightly under-estimate the required number n but the effect on n will be negligible if the resulting total sample size in each group is moderately large. The approach in which m is replaced by $m_{\max}$, the largest expected cluster size in the sample, provides some protection of statistical power in the event that the loss to follow-up rate for the trial is underestimated (Donner and Klar, 2000; Donner *et al*., 1981).

ii. Comparison of proportions

If the aim is to test the hypothesis $H_0 : P_1 = P_2$ at the two-sided 100α per cent level of significance with power $1-\beta$, where P_1 and P_2 are the success rates in the experimental and control groups, respectively, and $Z_{\alpha/2}$ and Z_{β} denote the critical values of the standard normal distribution corresponding to the error rates α and β, respectively then the required number of subjects per group, as given by Donner and Klar (2000: 62), is:

$$n \approx \frac{(Z_{\alpha/2} + Z_{\beta})^2 \left[\sum_{i=1}^{2} P_i(1-P_i)\right]\left(1 + [m-1]\rho\right)}{\left(P_1 - P_2\right)^2}, \tag{14}$$

and the number of clusters per group is approximated by $k = \dfrac{n}{m}$.

2.4.2 The matched–pair design

The matched-pair design is distinguished from the completely randomised and stratified designs in that it involves no replication of clusters within any combination of intervention and stratum. Estimates of random variability among clusters can only be disentangled from intervention effects by using between-stratum information (Donner and Klar, 2000).

i. Comparison of means

Suppose the primary response variable Y for the trial is normally distributed, and the matched-pair allocation yields the observed mean difference $d_j = \overline{Y}_{1j} - \overline{Y}_{2j}$, where $\overline{Y}_{1j}$ and $\overline{Y}_{2j}$ are computed over clusters of size m_{1j} and m_{2j}, respectively; and $j = 1,2,...,k$. Suppose the aim of the investigation is to determine the number of pairs k needed to test $H_0 : \Delta = 0$, given that Δ is the population mean of the d_j.

If we assume that $m_{1j} = m_{2j} = m$, $j = 1,2,...,k$ then the number of pairs required to test H_0 at the two-sided 100α per cent level of significance with power $1-\beta = 0.80$ is:

$$k = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 Var(d_j)}{\Delta^2} \qquad \textbf{(15)}$$

where, the variance of d_j is:

$Var(d_j) = 2\left(\frac{\sigma^2{}_W}{m} + \sigma^2{}_{AM}\right)$, with $\sigma^2{}_W$ denoting the within-cluster component of variability and $\sigma^2{}_{AM}$ denoting the component of variability between the two clusters in a matched pair (Donner and Klar, 2000: 66).

ii. Comparison of proportions

We denote the sizes of the clusters randomised to the experimental and control interventions as m_{1j} and m_{2j}, respectively; with $j = 1,2,...,k$. Let the observed sample differences be $d_j = \hat{P}_{1j} - \hat{P}_{2j}$ where $\hat{P}_{1j}$ and $\hat{P}_{2j}$ estimate the true event rates P_{1j} and P_{2j}, respectively.

If we are to determine the number of pairs, k, required to test $H_0 : \Delta = 0$, with Δ being the population mean of the d_j, then under the assumption that a paired t-test is to be performed, the required number of pairs needed to achieve power $1-\beta$ for detecting a difference Δ at two-sided 100α per cent significance level is:

$$k = \left(Z_{\frac{\alpha}{2}} + Z_{\beta} \right)^2 \frac{Var(d_j)}{\Delta^2}, \tag{16}$$

where $m_{1j} = m_{2j} = m$ and $Var(d_j) = \sum_{i=1}^{2} \frac{P_{ij}(1-P_{ij})}{m_{ij}} + 2\sigma^2{}_{AM}$, with $\sigma^2{}_{AM}$ representing the between-cluster variance component (Donner and Klar, 2000).

2.4.3 The stratified design

The stratified design is an extension of the matched-pair designs in which several clusters are randomly assigned within strata to each of the intervention and control groups (Donner and Klar, 2000: 33). The choice of stratification variables may be dictated as much by practical as by scientific considerations. Stratification is particularly important in smaller trials, where the probability of imbalance on important baseline factors can be substantially greater than in a trial randomising a comparable number of individuals (Donner, 1992).

i. Comparison of means

The stratified design can be regarded as a replication of the completely randomised design as implemented in each of k stratum, where the strata are defined by cluster-level variables measured at baseline (Donner and Klar, 2000: 70). The main impact of stratification on the assessment of sample size for testing the equality of means with respect to a quantitative response variable is obtained through the corresponding reduction obtained in the between-cluster variance component of the overall response variance σ^2. The actual gain in precision resulting from stratification is often difficult to judge in advance, however this gain will always be positive if stratification is effective and the degrees of freedom are large (Donner and Klar, 2000).

ii. Comparison of proportions

Woolson *et al* (1986) have presented sample size requirements for stratified analyses involving a set of 2 x 2 contingency tables in which only one of the margins in each table (stratum) is fixed. The formula they gave assumed that a Cochran statistic would be used to test the hypothesis that the assumed common odds ratio differs from unity and that the total number of individuals in each stratum is large (Donner, 1992).

Suppose that n_i clusters of size m_i are randomly assigned to the intervention and control groups in stratum $i, i = 1,2,...,k$. Let the success probability characterizing control group subjects in stratum $i, i = 1,2,...,k$, be P_{iC}, with P_{iT} denoting the probability of success among subjects in the treatment group. If we assume that the value of the intervention odds ratio

$$\psi = \frac{P_{iT}(1 - P_{iC})}{P_{iC}(1 - P_{iT})} \qquad \textbf{(17)}$$

is constant across strata, then (see Donner, 1992; Donner and Klar, 2000: 71) P_{iT} can be written as:

$$P_{iT} = \frac{P_{iC}\psi}{1 - P_{iC} + P_{iC}\psi}. \qquad \textbf{18)}$$

Let $\overline{P}_i = \frac{1}{2}(P_{iT} + P_{iC})$ denote the overall success rate for stratum i, and let $\sum_{i=1}^{k} t_i = 1$, where t_i denotes the fraction of individuals in the trial belonging to stratum i.

The total number of subjects required for testing $H_0 : \psi = 1$ is given (see Donner , 1992: 744; Donner and Klar, 2000: 71) by:

$$N_W = \frac{\left[Z_\alpha T + Z_\beta U\right]^2}{V^2} \tag{19}$$

where

$$T = \frac{1}{2}\sqrt{\left\{\sum_{i=1}^{k} t_i\left[\overline{P}_i\left(1-\overline{P}_i\right)\right]\right\}} \tag{20}$$

$$U = \sqrt{\left\{\frac{1}{8}\left[\sum_{i=1}^{k} t_i\left[P_{iT}\left(1-P_{iT}\right)+P_{iC}\left(1-P_{iC}\right)\right]\right]\right\}} \tag{21}$$

$$V = \frac{1}{4}\sum_{i=1}^{k} t_i\left(P_{iT}-P_{iC}\right) \tag{22}$$

and Z_α, Z_β denote the critical values of the standard normal distribution corresponding to the error rates α and β.

2.5 The unit of analysis

One of the most important ideas and/or aspects in any research project is what is regarded as the major entity that a researcher is analyzing, the unit of analysis, because this has an impact on the analysis to be carried out.

Therefore, the unit of analysis is the basic entity or object about which generalisations are to be made based on an analysis, and for which data have been collected. The unit of analysis is fundamental to data and statistical inference, to statistical data structures, and to secondary data analysis.

The unit of analysis can be artifacts, areas, groups, individuals or social interactions such as divorces, marriages, etc. In choosing the unit of analysis the researcher should be aware of the possibility of the following fallacies, or errors of reasoning based on mistaken assumptions:

a. The *ecological fallacy*, also called '*Robinson effect*', which occurs when one makes conclusions about individuals based only on the analyses of group data. That is, interpreting aggregated data at the individual level (Hox, 1995).
b. An *exception fallacy* occurs when one reaches a cluster/group conclusion on the basis of exceptional cases.
c. *Atomistic fallacy*, an error committed when one draws inference at a higher level from analyses performed at a lower level.

The unit of inference in cluster randomisation trials may be directed either at the cluster level or at the level of the individual subject, and an interest in cluster-level inferences leads investigators to collect data only at the cluster level (Donner and Klar, 2000). The investigator faced with analysing individual-level data must account for the lack of statistical independence among observations within a cluster.

Donner and Klar (2000: 81) suggest as a method of simplifying the problem, '*to collapse the data in each cluster, followed by the construction of a meaningful summary measure to serve as the unit of analysis'* so that the standard statistical methods can be directly applied to the collapsed measures. Challenges in identifying and distinguishing the unit of inference, the unit of analysis and the impact of clustering are not unique to cluster randomisation trials, since Donner and Klar (2000) reported that many of the ideas presented in cluster randomisation trials are based on discussions of methods for the analysis of longitudinal or repeated-measures data.

Cluster-level analyses can be used for more complex summary scores and, more generally, for any study outcome. These analyses (cluster-level) are most obviously appropriate when the primary questions of interest are more concerned with the randomised unit as a whole, than with the individual cases/subjects.

Donner and Klar (1994) emphasize the fact that the primary advantage of using the cluster as the unit of analysis is that standard techniques are generally applicable for any one of the three designs. There are, however, disadvantages to this approach.

Tests of significance using the cluster as the unit of analysis will, in general, have less power than methods using the individual as the unit of analysis. That is, the statistical power of cluster-randomised trials is greatly reduced in comparison with similar sized individually randomised trial (Campbell and Grimshaw, 1998). However, Donner and Klar (1994) quoting Shirley and Hickling (1981) point out that simulation studies have demonstrated that the loss in power is small, and using weighted least squares could increase the efficiency of these procedures.

Questions concerning the appropriate unit of analysis are more challenging when the primary target of inference is at the level of the individual subjects, with the choice of randomisation unit then largely a matter of convenience or other practical considerations. Particular care must be taken, according to Donner and Klar (2000: 81), when conducting individual-level analysis to properly adjust for the effect of clustering. Donner and Klar (2000) summarised the advantages of cluster-level and individual-level analyses, as shown in the table below.

Table 1: Comparing cluster-level and individual-level analyses

Cluster-level analyses	***Individual-level analyses***
• Easy to conduct and explain	• Reduce to the standard statistical procedures in the absence of clustering
• Can be applied to any outcome variable	• Allow more direct examination of the joint effects of cluster-level and individual-level predictors
• Permit the construction of exact statistical inferences	• Can be extended to permit analyses of multilevel data
• Can be adapted to adjust for baseline imbalance and/or variability in cluster size	• More naturally yield estimates of intracluster correlation coefficient
• When properly weighted will provide power comparable to individual-level analyses	• Will provide more efficient estimates of the effect of intervention than unweighted analyses when there are many cluster per group, particularly when cluster sizes are highly variable

2.6 Analysis of continuous outcomes

2.6.1 Aggregate model

The results obtained from a standard statistical analysis, which assumes all observations to be independent in clustered data, may be misleading. Such analysis is referred to as *naïve pooling* (Burton *et al*., 1998). Therefore, standard analysis, which ignores an important correlation structure, may well be misleading. One way of solving this problem is to create a single summary statistic for each cluster, for example the mean, and this approach, which may be referred to as *data resolution*, automatically avoids any over-inflation in the apparent size of the data set (Burton *et al., 1998)*. The best statistic to use will depend upon the research question being asked.

There are sometimes occasions when the only data available for analysis have already been aggregated to a higher level. For example, as stated by Goldstein (1995: 63), we may have information on student achievement only in terms of the mean achievement for each school, or information on utilization of health services only in terms of total number of episodes for each administrative area. According to Wears (2000) this is the simplest alternative to individual analysis in that one simply reduces all the individual observations within a cluster to a single summary measure, such as cluster mean or proportion, and then use standard statistical methods to analyze these summary measures as if they are primary measures.

In comparison of these *summary statistics* between the treatment group and control group, the standard independent t-test can be employed.

There is nothing wrong with aggregation in cases where the researcher is only interested in macro-level propositions, although it should be borne in mind that the reliability of an aggregated variable depends, among other things, on the number of micro-level units in a macro-level unit, and thus will be larger for the larger macro-units than for smaller ones (Snijders and Bosker, 1999). Snijders and Bosker (1999: 09) give explicit definitions and/or explanations of macro-level, micro-level, and cross-level relations.

This approach, aggregation, has the advantage of not requiring specialised software, but usually entails a drastic reduction in sample size and degrees of freedom, and consequently statistical power (Wears, 2000). This approach also suffers from vulnerability to the *ecological fallacy*. Here, an analyst must be aware of '*Simpson's Paradox*' which refers to the situation in which completely erroneous conclusions may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogeneous population (Hox, 1995).

Simpson's paradox is a statistical paradox described by E. H. Simpson (1951), in which the successes of several groups seem to be reversed when the groups are combined, and this seemingly impossible result is encountered surprisingly often in social science and medical statistics (Wikipedia). According to Pearl (2000) Simpson's paradox (Simpson, 1951; Blyth, 1972), first encountered by Pearson in 1899 (Aldrich, 1995), refers to the phenomenon whereby an event C increases the probability of E in a given population p and, at the same time, decreases the probability of E in every subpopulation of p. Spilker (1991: 544) gives a similar explanation of Simpson's paradox with examples.

2.6.2 Mixed model

Analysis of continuous outcome data from cluster randomisation trials can often be accomplished using mixed-effects linear models, as fitted by generalised least squares (Donner, 1985). Lin *et al*. (1997) point out that linear mixed models have been widely used in the analysis of clustered and longitudinal data. And in such models one of the common assumptions is that the within-cluster variances are homogeneous. The mixed-effects linear models can be used with stratified or completely randomised designs to estimate the effect of treatment, to test if the observed effect is due to chance, and to adjust for imbalance on baseline risk factors (Donner, 1994).

A mixed-effect model as given by (Donner, 1994; Donner and Klar, 2000) is:

$$y_{ijst} = \beta X_{ijst} + C_{(ij)s} + \varepsilon_{(ijs)t}, \qquad \textbf{(23)}$$

where Y_{ijst} is the score for the t[th] subject, $t = 1,...,n_{ijs}$, from cluster s, $s = 1,...,m_{ij}$, treatment $j, j = 1,2$ and strata $i, i = 1,...,k$. The vector β represents the fixed effect of treatment, strata and baseline risk factors. $C_{(ij)s}$ and $\varepsilon_{(ijs)t}$ denote the respective independent random effects of cluster nested in treatment and stratum, assumed to be $i.i.d.$ $N(0, \sigma^2{}_C)$, and subject nested in cluster, assumed to be $i.i.d. N(0, \sigma^2)$.

A special case of the mixed effects linear regression model is an approach, which is based on a two-stage nested analysis, for testing the hypothesis $H_o : \mu_1 = \mu_2$, where μ_1 and μ_2 are the underlying means of the outcome variable y. This model is given by

$$y_{ijk} = \mu + \beta_1 x_{1i} + u_{ij} + \varepsilon_{ijk} \qquad \textbf{(24)}$$

where the term μ in the model is the true mean response (grand mean), and β_1 is a regression constant representing the fixed effect of intervention group $i(i = 0,1)$ such that

$$x_{1i} = \begin{cases} 1(Active\_Arm), i = 1 \\ 0(Control\_Arm), i = 0 \end{cases},$$

and u_{ij}, which is assumed to be normally distributed with mean 0 and variance $\sigma^2{}_A$, denotes a random cluster effects. Also we assume an error term ε_{ijk} is normally distributed with mean 0 and variance $\sigma^2{}_W$.

The effects u_{ij} and ε_{ijk} are assumed to be mutually independent, with a common intracluater correlation given by $\rho = \dfrac{\sigma^2{}_A}{\sigma^2{}_A + \sigma^2{}_W}$, showing that this parameter measures the proportion of variance in y_{ijk} that can be attributed to the variance among clusters (Donner and Klar, 2000).

A summary analysis of variance (ANOVA) table, which is strictly applicable for balanced designs, also which may be used to test the effect of intervention, i.e. to test $H_0 : \beta_1 = 0$, $i = 0,1$ is given by Donner and Klar (2000: 117), as follows:

Table 2: ANOVA table for testing the intervention effect

Source of variation	Degree of freedom	Sum of squares	Mean square	F
Among intervention groups	*1*	*SSG*	*MSG*	MSG/MSC
Cluster within groups	*2(k-1)*	*SSC*	*MSC*	
Subjects within clusters	*2k(m-1)*	*SSW*	*MSW*	

If ρ is erroneously assumed to be zero, equivalent to omitting the terms u_{ij} from the model, the resulting analysis will have an inflated type I error (Donner and Klar, 2000).

Covariate adjustment is more powerful than an analysis based on change scores, particularly when the correlation between the baseline covariate and response variable is weak. The regression approach also allows the intervention comparison to be adjusted for the influence of several baseline covariates simultaneously, measured at either the cluster or individual level (Donner and Klar, 2000: 121). Therefore equation (**18**) can be re-expressed as:

$$y_{ijk} = \mu + \beta_1 x_{1ijk} + \beta_2 x_{2i} + u_{ij} + \varepsilon_{ijk} \tag{25}$$

where

β_1 measures "covariate effect",

β_2 measures "treatment effect", and

$$x_{2i} = \begin{cases} 1, active\_arm \\ 0, control\_arm \end{cases}$$

Multi-level models with complex variation allow the treatment effect to vary from cluster to cluster, and equation (**25**) is re-expressed as:

$$y_{ijk} = \mu + \beta_1 x_{1ijk} + (\beta_2 + v_{ij}) x_{2i} + u_{ij} + \varepsilon_{ijk} \tag{26}$$

where v_{ij} measures the treatment effect that varies from cluster to cluster and is assumed to be normally distributed with mean 0 and variance $\sigma^2{}_V$ i.e. $v_{ij} \sim N(0, \sigma^2{}_v)$.

Such adjustments can further increase the precision with which intervention effects are estimated, as well as account for chance imbalance on those variables that may distort the estimated effect of the intervention. Complex multilevel models allow, for example, the effect of treatment to vary randomly from cluster to cluster.

The matrix notation, with detailed explanation and assumptions, for mixed models is given by, amongst others, Longford (1993), Hox (1995), Davidian and Giltinan (1996), and Snijders and Bosker (1999).

In general, it is worthwhile to adjust for those covariates that simultaneously influence outcome and have different distributions in the two intervention groups (Hauk *et al.*, 1998). Nonetheless, covariate adjustment for a factor that is strongly related to outcome can improve precision even though it has the same distribution in both groups.

Conversely, adjustment for a weakly correlated covariate may dilute statistical power and even compromise the validity of statistical inference when there are a limited number of clusters per intervention group (Donner and Klar, 2000).

An alternative way of looking at data with two variance components is referred to in the behavioural science and educational literature as hierarchical linear modelling or multilevel modelling (Goldstein, 1995).

2.6.3 Multilevel modeling or hierarchical modeling

Medical research applications often involve hierarchical or multilevel data structures or nested data structures such as patients within hospitals or physicians within hospitals; for example, assessing differences in mortality rates across hospitals relative to a specific condition or procedure (Sullivan *et al.*, 1999).

Multi-level modelling has been developed as a technique for analysing data arranged in a variety of hierarchies or clusters (Langford *et al.*, 1998; Hox, 1995), that is multilevel analysis is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of variability (Snijders and Bosker, 1999). According to Snijders and Bosker (1999: 1) multilevel analysis is a stream that has two tributaries:

a. *Contextual analysis*, which was developed in social sciences, focused on the effect of the social context on individual behaviour, and
b. *Mixed effect models*, which are statistical models in the analysis of variance and in regression analysis where it is assumed that some of the coefficients are fixed and others are random.

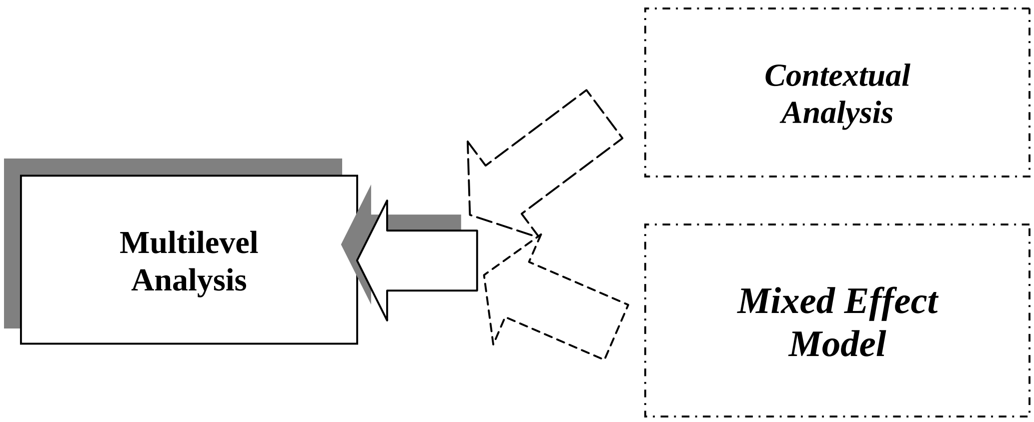

Figure 5: Multilevel analysis

Contextual modeling until about 1980 focused on the definition of appropriate variables to be used in ordinary least squares regression analysis. The main focus in the development of statistical procedures for mixed models up to the1980s was on random effects (i.e., random differences between classes in some classification system) rather than on random coefficients (i.e., random effects of numerical variables). The two streams, mixed models and contextual modeling came together to form multilevel analysis.

Nested, or hierarchical, data present problems for analysis. People or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population (Osborne, 2000).

In investigating the relationship between study subjects and society, generally individuals interact with the social contexts to which they belong, meaning that individual persons are influenced by the social groups to which they belong, and that the properties of those groups are in turn influenced by the individuals who make up that group (Hox, 1995).

Goldstein (1995: 2) stated a well known and influential study of primary school children carried out in the 1970s (Bennette, 1976), which claimed that children exposed to so-called ‘formal’ styles of teaching exhibited more progress than those who were not. The data were analysed using traditional multiple regression techniques which recognised only the individual children as the units of analysis and ignored their groupings within teachers and into classes.

The results were statistically significant. Subsequently, Aitkin et al (1981) demonstrated that when the analysis accounted properly for the grouping of children into their classes, the significant differences disappeared and the ‘formally’ taught children could not be shown to differ from the others. This re-analysis was the first important example of a multilevel analysis of social science data.

A hierarchy consists of units grouped at different levels. Thus learners may be the level-1 units in a 2-level structure where the level 2 units are the schools. The following diagram can represent an example of hierarchical data.

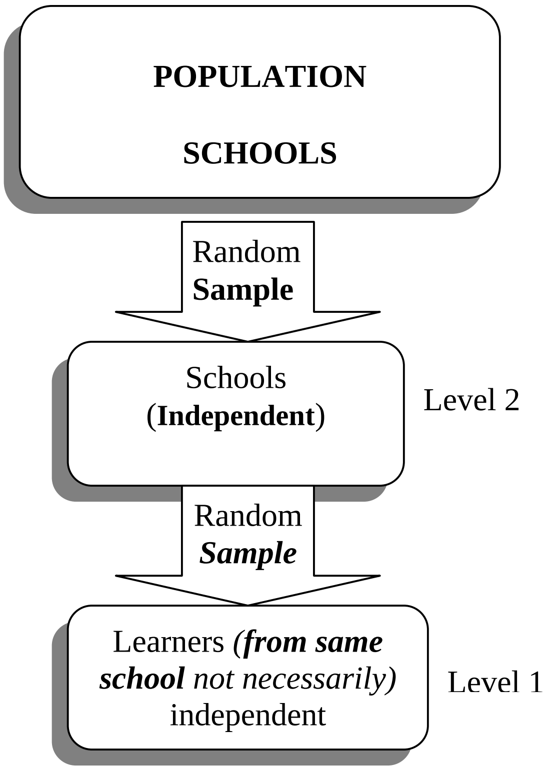

Figure 6: Example of population with hierarchical structure

In multilevel research, variables can be defined at any level of the hierarchy, and some of these variables may be measured directly at their natural level (Hox, 1995). In these schemes, as represented in Figure 6, individuals usually form micro-level (lowest level or level 1) whereas groups form macro-level (level 2).

A more general way to look at multilevel data is to investigate a cross level hypothesis, *or multilevel problem* (Hox, 1995). A multilevel problem concerns the relationships between variables that are measured at a number of different hierarchical levels. Thus, multilevel models are designed to analyse variables from different levels simultaneously, using a statistical model that includes the various dependencies. Multilevel statistical models are always needed if a multi-stage sampling design has been employed (Levy and Lemeshow, 1999; Snijders and Bosker, 1999).

The main statistical model of multilevel analysis is the hierarchical linear model, an extension of the multiple linear regression-model to a model that includes nested random coefficients. The multilevel regression model is also known as random coefficient model, variance component model, or hierarchical linear model (Hox, 1995).

The full multilevel regression model assumes that there is a hierarchical data set, with one single dependent variable that is measured at the lowest level and explanatory variables at all existing levels (Hox, 1995).

Hierarchical linear models, sometimes called multi-level linear models, nested models, mixed linear models or covariance component models, handle hierarchical data structures, and these models have historically been used in educational research where hierarchies occur naturally (Sullivan *et al*., 1999).

According to Donner and Klar (2000:118) investigators have recognized that there is a very close relationship between mixed effect linear regression and multilevel modeling.

Chapter 3

Possible analyses for binary (discrete) outcomes

Methods for analyzing binary outcome data in cluster randomisation trials are not as well established as methods for analyzing continuous outcome data (Donner and Klar, 2000; Donner and Klar, 1994). The most obvious reason is that linear models were developed for continuous data earlier than the nonlinear models generally used for binary data (Donner and Klar, 1994).

The analysis of binary data from cluster randomisation studies poses greater challenges than that of continuous outcomes. Because all the conventional methods are approximate, and there is no universally superior approach (Ukoumunne and Thompson, 2001) i.e., virtually all the methods that can be developed are approximate; there is no unique multivariate extension of the binomial distribution (Donner, 1998).

Until recently, appropriate analyses of cluster randomised trials with binary outcomes were mostly performed either using cluster-level summary statistics, with methods such as the t-test, or by adjusting standard methods such as the chi-squared test to allow for the effects of clustering (Turner *et al*., 2001). The most accessible or simplest approach to analyzing binary data from a cluster randomisation trial is to obtain a single summary score for each cluster, so that the analysis can be conducted at the same level as the random assignment using standard statistical procedures (Donner and Klar, 2000).

3.1 Completely randomised design

3.1.1 Comparison of proportions

The principal objective of many cluster randomisation designs is to compare the proportion of individuals in different treatment groups who have a specific characteristic. The case of a two-group intervention study will be considered, since the procedures can be generalised in a natural way to compare more than two groups.

Suppose that k_i clusters are randomised to intervention group $i(i = 1,2)$,

m_{ij} being the size of the j^{th} cluster in group i, and

Y_{ijl}, $l = 1,...,m_{ij}$, be the binary outcome for the l^{th} subject from the ij^{th} cluster.

Let $Y_{ijl} = 1$ for subjects with outcome of interest (success) otherwise $Y_{ijl} = 0$. Therefore,

$Y_{ij} = \sum_{l=1}^{m_{ij}} Y_{ijl}$ denotes the corresponding number of subjects with outcome of interest, and

$\hat{P}_{ij} = \dfrac{Y_{ij}}{m_{ij}}$ denotes the cluster-specific event rate or the proportion of subjects in the cluster who develop the specific characteristic over the length of study, $j = 1,2,3,...,k_i$.

Let us define the following notations:

$Y_i = \sum_{j=1}^{k_i} Y_{ij}$ (Total number of successes in the i^{th} *group*)

$M_i = \sum_{j=1}^{k_i} m_{ij}$ (Total number of individuals in the i^{th} group)

$K = \sum_{i=1}^{2} k_i$ (Total number of clusters)

$M = \sum_{i=1}^{2} M_i$ (Total number of individuals)

$Y = \sum_{i=1}^{2} Y_i$ (Total number of successes)

$\hat{P}_i = \dfrac{Y_i}{M_i}$ (Event rate as computed over all clusters in i^{th} group)

$\hat{P} = \dfrac{Y}{M}$ (Overall event rate observed in the study)

Let ρ be the intracluster correlation coefficient (assuming a common value). A pooled estimate of this parameter is obtained by applying the standard analysis of variance formula to the (0,1) binary observations (Donner and Klar, 2000). The estimator of ρ may be written as:

$$\hat{\rho} = \frac{MSC - MSW}{MSC + (m_0 - 1)MSW} \qquad \textbf{(27)}$$

where

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

$$MSW = \frac{\sum\sum m_{ij}\hat{P}_{ij}(1-\hat{P}_{ij})}{M-K}$$

with

$$m_0 = \frac{M - \sum_{i=1}^{2}\overline{m}_{Ai}}{K-2} \text{ and } \overline{m}_{Ai} = \sum_{j=1}^{k_i} m^2{}_{ij} / M_i \,.$$

MSC and MSW are the pooled mean square errors between and within clusters, respectively. The (null) hypothesis, $H_0 : P_1 = P_2$, that the proportions in the two intervention groups are not statistically different will be tested.

3.1.2 Cluster-level analysis

a) Two-sample t-test

The two-sample t-test is employed in the comparison of average values of the event rate in the two groups. Let the unweighted mean event rate be denoted by

$$\overline{P}_i = \sum_{j=1}^{k_i} \hat{P}_{ij} \big/ k_i \,, \qquad \boldsymbol{i = 1, 2.} \qquad \textbf{(28)}$$

Also denote the sample variance of the observed event rate in the groups by $S_1^{\ 2}$ and $S_2^{\ 2}$, respectively, where

$$S_i^{\ 2} = \sum_{j=1}^{k_i} \frac{\left(\hat{P}_{ij} - \overline{P}_i\right)^2}{k_i - 1} \,, \qquad i = 1,2$$

Therefore, the unpaired two-sample (*robust*) t-statistic is given by:

$$t = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\left[\frac{\sum_{i=1}^{2}(k_i - 1)S_i^{\,2}}{K-2}\right]\left[\sum_{i=1}^{2}\frac{1}{k_i}\right]}} . \qquad \textbf{(29)}$$

Under the null hypothesis of no difference between the true mean event rates, *t* follows an approximate *t*-distribution with $K - 2$ degrees of freedom (Donner and Klar, 2000).

b) Non-parametric approach

A non-parametric approach**,** such as the Wilcoxon rank sum test, also known as the Mann-Whitney U-test, is called upon when the parametric t-test is not applicable, i.e. the t-test assumes that the data are normally distributed, and if this assumption is violated then a non-parametric approach becomes applicable. This approach is based on the ranks of observed event rates. This procedure requires that the two groups be pooled and the event rates ranked by size (Donner and Klar, 2000).

If the intervention has no effect, then the sum of ranks for each group should be the same. Let the number of clusters be at least (15) fifteen, and denote the sum of ranks for the ith group by $T_i \ \ (i = 1,2)$. According to Donner and Klar (2000: 88), one performs a test by referring

$$Z_w = \frac{T_1 - \frac{k_1(k_1 + k_2 + 1)}{2}}{\sqrt{\frac{k_1 k_2 (k_1 + k_2 + 1)}{12}}} \qquad \textbf{(30)}$$

to tables of the standard normal distribution. It is noted that inference procedures based on ranks ignore not only variations in cluster size, but also the actual magnitudes of the event rates (Conover, 1980).

c) Fisher's two-sample permutation

Fisher's two-sample permutation test takes into account the values of the cluster-specific event rates (Conover, 1980: 327). The two-sided statistical significance of the test equals the proportion of test statistics found using the permuted data that are at least as large, in absolute value, as the test statistic found using the observed data (Donner and Klar, 2000). The estimated standard error of $\bar{P}_1 - \bar{P}_2$, which is an approximate version of the permutation test, is given by:

$$\hat{S}_e(\bar{P}_1 - \bar{P}_2) = \left\{ \frac{\sum_{i=1}^{2}\sum_{j=1}^{k_i}\left(\hat{P}_{ij} - \bar{P}\right)^2}{K-1} \left(\frac{1}{k_1} + \frac{1}{k_2}\right) \right\}^{\frac{1}{2}} \qquad (31)$$

where

$$\bar{P} = \sum_{i=1}^{2}\sum_{j=1}^{k_i} \hat{P}_{ij} \big/ K$$

3.1.3 Individual-level analyses

One individual-level approach for testing the difference between event rates is to empirically adjust the standard chi-square test for the clustering of responses within a randomised unit. This method yields the standard Pearson chi-square statistic if the estimated intra-cluster coefficient $\hat{\rho}$ is zero. It can be extended to the comparison of more than two groups with ease.

a) Adjusted chi-square approach

The adjusted chi-square approach as proposed by Donner and Donald (1988) depends on computing a clustering correction factor for each group, and the correction factor is given by:

$$C_i = \frac{\sum_{j=1}^{k_i} m_{ij}\left[1 + \left(m_{ij} - 1\right)\hat{\rho}\right]}{\sum_{j=1}^{k_i} m_{ij}} = 1 + \left(\bar{m}_{Ai} - 1\right)\hat{\rho} \qquad i = 1,2 \qquad (32)$$

The factor C_i can be thought of as the estimated or observed design effect in group i, where the population design effect measures the inflation in variance of $\hat{P}_i$ that can be attributed to clustering (Donner and Klar, 2000). It is assumed, under this adjustment, that factor 1 and factor 2 estimate the same population design effect. This assumption is guaranteed to hold for experimental comparisons, and may not hold for non-randomised (or observational) comparisons, in which clusters are systematically assigned to receive different interventions (Donner *et al*, 1994). The adjusted one-degree of freedom chi-square statistic as quoted by Donner and Klar (2000: 90) is:

$$\chi_A{}^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i \hat{P}\left(1 - \hat{P}\right)} \qquad \textbf{(33)}$$

It is important to note that the validity of the adjusted chi-square test statistic does not depend on the assumption that the pair-wise correlation between any two observations in the same cluster is constant, i.e. it does not require an 'exchangeable' correlation structure (Donner and Klar, 2000). In the case of a heterogeneous correlation structure, $\hat{\rho}$ measures the average degree of pair-wise correlation that is assumed to be constant across clusters.

b) Ratio estimator approach

The ratio estimator approach is also based on a relatively simple adjustment of the standard Pearson chi-square statistic. If individuals rather than clusters were the unit of allocation, then the estimated variance of $\hat{P}$ would be given by the binomial expression

$$\hat{V}_B(\hat{P}_i) = \frac{\hat{P}_i(1 - \hat{P}_i)}{M_i}, \qquad \textbf{(34)}$$

which tends to underestimate the true variance of $\hat{P}_i$ if the unit of allocation is an intact cluster (Donner and Klar, 2000). If $\hat{P}_i$ is the ratio then, according to Cochran (1977), an appropriate variance estimate is given by:

$$\hat{V}_R(\hat{P}_i) = k_i(k_i - 1)^{-1} M_i{}^{-2} \sum_{j=1}^{k_i} \left(Y_{ij} - m_{ij}\hat{P}_i\right)^2. \qquad \textbf{(35)}$$

The estimated design effect in group i is given by $d_i = \dfrac{\hat{V}_R(\hat{P}_i)}{\hat{V}_B(\hat{P}_i)}$, and these estimates may be used to adjust the standard Pearson chi-square statistic for the effect of clustering (Donner and Klar, 2000). Rao and Scott (1992), as quoted by Donner and Klar (2000), gave the one-degree of freedom estimator of chi-square statistic as:

$$\chi_R^{\ 2} = \sum_{i=1}^{2} \frac{\tilde{M}_i(\hat{P}_i - \tilde{P})}{\tilde{P}(1-\tilde{P})} \qquad \textbf{(36)}$$

where $\tilde{M}_i = {}^{M_i}\!/_{d_i}$, $\tilde{P} = \dfrac{\sum_{i=1}^{2}\tilde{Y}_i}{\sum_{i=1}^{2}\tilde{M}_i}$ and $\tilde{Y}_i = {}^{Y_i}\!/_{d_i}$.

The ratio estimator approach is applicable to non-randomised comparisons, particularly those involving systematic differences in mean cluster size from group to group.

c) Parametric modelling

In parametric modelling the subject responses are assumed to follow a specified probability distribution. That is, this approach is based on the adoption of the parametric model for characterizing the variability in the response among clusters (Donner *et al.* 1994). Several such models are available, depending on the scale chosen for the between-cluster variability and the underlying distribution such as the beta-binomial distribution, which results when the number of responses within a cluster follows a binomial distribution conditional on the probability parameter P_{ij}, and the cluster-specific success rates P_{ij} are assumed to vary across clusters in accordance with the beta distribution.

This implies that the resulting beta-binomial distribution will exhibit more variability than the binomial distribution but will essentially reduce to the afore-mentioned second assumption when the intra-cluster correlation coefficient is zero (Donner and Klar, 2000).

The parameters of the beta-binomial distribution may be estimated using the method of maximum likelihood estimation, which leads to a likelihood ratio test of the null hypothesis that a single beta-binomial distribution can be used to fit the observations (Donner and Klar, 2000: 92). The problems with likelihood ratio tests are that they are computationally intensive and very tied to parametric assumptions; also they may be unreliable in smaller samples and are not robust to departures from the assumed model (Donner *et al*., 1994).

d) Generalised Estimating Equation (GEE)

A 'population-averaged' method is one of three computationally intensive methods (Donner and Klar, 1994), and an example of the population-averaged method is Generalised Estimating Equations (GEE) introduced by Liang and Zeger (1986). Liang and Zeger (1986) introduced a class of estimating equations, based on moment methods, that provide consistent estimates of regression parameters and their variances in the context of generalised linear models that make no distributional assumptions (Donner *et al.,* 1994). The generalised estimating equations (GEE) approach of Liang and Zeger has become an established method for analysing correlated data (FitzGerald, 2002). The GEE is a non-linear analog of the marginal model, of a mixed-effect model defined in equation (**23**), described by:

$$Y_{ijs(t)} = \beta X_{ijst} + w_{(ijs)t}$$

where $w_{(ijs)t} = C_{(ij)s} + \varepsilon_{(ijs)t}$ is the t^{th} element of the vector w_{ijs}, each of which is *i.i.d.* multivariate normal random variables with a mean of zero and a variance matrix denoted by $\left(\sigma^2{}_C + \sigma\right)V$, with the V matrix being a correlation matrix in which all correlation coefficients are equal implying that responses of all cluster members are equally correlated (Donner and Klar, 1994).

This approach can also be used to construct an extension to standard logistic regression, which adjusts for the effect of clustering. In the absence of clustering, this approach reduces to that of standard logistic regression (Donner and Klar, 2000).

There are two strategies to adjust for the effect of clustering using this approach (GEE). The first can be said to be model-based, as it requires the specification of a working correlation matrix that describes the pattern of correlation between responses of cluster members, with the assumption that responses of cluster members are equally correlated, i.e. exchangeable (Donner and Klar, 2000).

The second strategy is to employ 'robust variance estimators' that are constructed using between-cluster information. If the variance form is mis-specified the parameter estimates in the model are still consistent, but not the corresponding variance estimator (Donner *et al.*, 1994). For large a number of clusters, inference obtained using robust variance estimation will become equivalent to those obtained using the model-based strategy when the working correlation matrix is correctly specified (Donner *et al.*, 1994; Donner and Klar, 2000).

An important advantage of the GEE approach is that it can readily be extended to adjust for a combination of cluster-level and individual-level covariates, as described by Moore and Tsiatis (1991), yielding consistent and asymptotically normally distributed estimators of regression coefficients (Donner *et al.*, 1994).

Generally, when the number of clusters is small and the number of individuals in each cluster is large, hypothesis tests may be based on Liang and Zeger's (1986) GEE approach, but standard GEE will not be appropriate in the sparse situation, where the number of clusters is moderately large, but the number of patients in many of the clusters is small (Zhang and Boos, 1996).

3.1.4 Confidence Intervals

As mentioned elsewhere cluster randomisation trials are frequently underpowered, and the interpretation of non-significant results from comparative trials is enhanced by their presentation in terms of confidence limits. According to Donner and Klar (2000: 96) an appropriate two-sided $(1-\alpha)100$ percent confidence interval for the effect of intervention estimated as a rate difference may be obtained as:

$$\left(\hat{P}_1 - \hat{P}_2\right) \pm Z_{\alpha/2}\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)C_2}{M_2} + \frac{\hat{P}_1(1-\hat{P}_1)C_1}{M_1}} \tag{37}$$

where $Z_{\alpha/2}$ is the $(1-\alpha)100$ per cent two-sided critical value of the standard normal distribution, and C_i are defined by **(32)** while $\hat{P}_i$ is the event rate as computed over all clusters in ith group.

A useful measure of risk in comparative studies is the odds ratio, ψ, defined by

$$\psi = \frac{P_1 Q_2}{P_2 Q_1}$$

where P_1 and P_2 are the prevalence in the two groups and $Q_i = 1 - P_i$ for $i = 1,2$ (Donald and Donner, 1987). Thus, the effect of the intervention may alternatively be expressed in terms of the odds ratio ψ, and the sample odds ratio $\hat{\psi}$ remains a consistent estimator of ψ in the presence of clustering (Donner and Klar, 2000), where

$$\hat{\psi} = \frac{\hat{P}_1(1-\hat{P}_2)}{\hat{P}_2(1-\hat{P}_1)}.$$

According to Donald and Donner (1987), the estimated variance of $\ell n(\hat{\psi})$ is given by

$$\hat{V}_O(\ell n[\hat{\psi}]) = \sum_{i=1}^{2} \frac{C_i}{M_i \hat{P}_i (1-\hat{P}_i)}.$$

Hence approximate two sided $(1-\alpha)100$ per cent confidence limits for ψ are given by:

$$\left(e^{\ell n\hat{\psi} - Z_{\alpha/2}\sqrt{\hat{V}_O(\ell n\hat{\psi})}}, e^{\ell n\hat{\psi} + Z_{\alpha/2}\sqrt{\hat{V}_O(\ell n\hat{\psi})}} \right).$$

3.2 Analysis of the matched-pair design

3.2.1 Comparison of proportions

a) Mantel and Haenszel chi-square test

Mantel and Haenszel extended the odds ratio estimator and the chi-square test of $H_o : \psi = 1$ to situations in which the data are stratified into $K > 1$ tables (Donald and Donner, 1987). Donald and Donner (1987) gave the general representation for the K 2×2 from a matched-pair cluster randomisation trials with dichotomous outcome variable, and the general formula for Mantel-Haenszel chi-square statistic.

In particular, the data from a matched-pair cluster randomisation trial with a dichotomous outcome variable, according to Donner and Klar (2000: 100), may be summarised in a series of k 2×2 contingency tables of the form:

Table 3: Data arising from a matched-pair cluster randomisation design

	Experimental	**Control**	**Total**

Event	a_{1j}	a_{2j}	r_{1j}
No event	$m_{1j} - a_{1j}$	$m_{2j} - a_{2j}$	r_{2j}
Total	m_{1j}	m_{2j}	m_j

where $r_{1j} = a_{1j} + a_{2j}$, $j = 1,2,...,k$ and $m_j = m_{1j} + m_{2j}$ the total sample size for the j^{th} table.

The *Mantel-Haenszel one degree of freedom chi-square* statistic for testing the null hypothesis that $\psi = 1$ in a set of K 2×2 contingency tables is:

$$\chi^2_{MH} = \frac{\left\{\sum_{j=1}^{k}\left(a_{1j} - m_{1j}r_{1j}/m_j\right)\right\}^2}{\sum_{j=1}^{k}\left\{\frac{m_{1j}m_{2j}r_{1j}\left(m_j - r_{1j}\right)}{m^2{}_j\left(m_j - 1\right)}\right\}}. \tag{38}$$

b) Liang's test procedure

The tests proposed by Liang (1985) which handle the sparse correlated binary data proved not to be very efficient since the variance estimates in the tests use the cluster as the primary sampling unit and are thus based on *n* degrees of freedom (*n* is the total number of clusters). Liang (1985) and Liang *et al.* (1986) therefore developed methodology for analysing 2×2 contingency tables when the binomial assumptions are relaxed (Donner and Klar, 2000). Liang's one degree of freedom chi-square statistic is given by:

$$\chi_L{}^2 = \frac{\left\{\sum_{j=1}^{k}(a_{1j} - m_{1j}r_{1j}/m_j)\right\}^2}{\sum_{j=1}^{k}\left(a_{1j} - m_{1j}r_{1j}/m_j\right)^2} \tag{39}$$

Donald and Donner (1987) gave a statistic that requires an assumption of common correlation across the strata. This statistic, the adjusted Mantel-Haenszel test statistic for correlated binary data, is based on estimating the intra-class correlation. The estimator of intracluster correlation is not stable in sparse data (Zhang and Boos, 1996).

Simulation studies have shown that Mantel-Haenszel is far too liberal when $\rho > 0$ (Song and Ahn, 2003). Liang's test tends to be conservative when the number of clusters is small, but approximates the 5% level well for large number of clusters (fifteen to twenty-five), and the Donald-Donner statistic has erratic behaviour for large number of clusters (Zhang and Boos, 1996). That is, Liang's test is over-conservative, but it behaves normally when the number of clusters is between fifteen to twenty-five.

c) Dependent means t-test

The dependent means t-test, also known as matched-pairs or paired-samples t-test, is used when there are two experimental conditions and the same subjects took part in both conditions of the experiment (Field, 2000: 207). With paired data we are interested in the average difference between the observations for each individual and the variability of these differences. We are thus interested in the variability of the within-subject differences rather than between-subject variation (Altman, 1991).

A parametric approach is obtained by applying this test to the differences d_j (Donner and Klar, 2000). The following table demonstrates the paired-sample t-test set-up.

Table 4: Comparing the two conditions (Repeated Measurements)

Index j	**Condition A** x_{mj}	**Condition B** y_{mj}	**Difference** $d_j = x_{mj} - y_{mj}$
1	x_{m1}	y_{m1}	$d_1 = x_{m1} - y_{m1}$
2	x_{m2}	y_{m2}	$d_2 = x_{m2} - y_{m2}$
.	.	.	.
.	.	.	.
.	.	.	.
k	x_{mk}	y_{mk}	$d_k = x_{mk} - y_{mk}$

The paired-sample t-test does not have the normality and equality of variances assumption, but assumes instead that the differences, d_j, come from a normally distributed population of differences (Zar, 1996). The paired t-statistic is given by:

$$t_p = \frac{\bar{d}\sqrt{k}}{\sqrt{\sum_{j=1}^{k} \frac{\left(d_j - \bar{d}\right)^2}{k-1}}} \quad ; \quad \bar{d} = \sum_{j=1}^{k} d_j \Big/ k$$

with $(k-1)$ degree of freedom, and it is fairly *robust* to departures from these assumptions.

d) Non-parametric approaches

Non-parametric approaches could also be used to test the null hypothesis, H_0, that two treatments conditions are not statistically different. One such an approach is the Wilcoxon Signed-rank test. The Wilcoxon test, also known as Wilcoxon matched pairs test, is used in situations in which there are two sets of scores to compare, but these scores come from the same subject (Field, 2000).

This test is often regarded as being similar to a matched-pairs t-test, and it is used to determine differences between groups of paired data when the data do not meet the rigor associated with a parametric test (MacFarland, 1998), i.e. one wants to see whether the populations from which the pairs of observations are drawn are the same. Therefore, if the population is compatible with the specified central point, then not only should there be about the same number of observations above and below the median, but the ranked observations should be more or less symmetrically distributed about the median.

According to Donner and Klar (2000) non-parametric approaches avoid distributional assumptions, but as a consequence suffer from reduced power, and they do not take into account the variable cluster sizes that characterize most trials. The Wilcoxon signed rank test is applied to the $d_j = \hat{P}_{1j} - \hat{P}_{2j}$ based on the ranks of the d_j rather than on their actual values. Details on how to carry out the Wilcoxon test are given by Mendenhall *et al.* (1981).

Clusters having identical absolute values of d_j are assigned the mean ranks. If T denote the sum of ranks for those clusters having $d_j > 0$, the test procedure is performed by comparing the value of T with the corresponding sum of ranks expected under H_0. According to Mendenhall *et al.* (1981: 585) T will be approximately normally distributed when the null hypothesis is true and k (the number of strata) is large, and it is for this reason that one is able to construct a large-sample Z test, where

$$E(T) = \frac{k(k+1)}{4} \text{ and } Var(T) = \frac{k(k+1)(2k+1)}{24}$$

then the Z statistic,

$$Z_{WS} = \frac{T - E(T)}{\sqrt{Var(T)}} = \frac{T - \left[\frac{k(k+1)}{4}\right]}{\sqrt{\frac{k(k+1)(2k+1)}{24}}} \tag{40}$$

can be used as a test statistics.

3.2.2 Confidence Intervals

The Mantel-Haenszel chi-square statistic is not valid for matched-pair cluster randomisation designs, but the associated Mantel-Haenszel estimator $\hat{\psi}_{MH}$ of an assumed common odds ratio ψ remains consistent (Donner and Klar, 2000). The Mantel-Haenszel estimator is given by:

$$\hat{\psi}_{MH} = \frac{\sum_{j=1}^{k} a_{1j}(m_{2j} - a_{2j})/m_j}{\sum_{j=1}^{k} a_{2j}(m_{1j} - a_{1j})/m_j} = \frac{\sum_{j=1}^{k} U_j}{\sum_{j=1}^{k} V_j} \tag{41}$$

For fairly large number of pairs (strata), the large sample variance of $\hat{\psi}_{MH}$ may be estimated from between-stratum information, and is given by:

$$\hat{V}(\hat{\psi}_{MH}) = \frac{\sum_{j=1}^{k}(U_j - \hat{\psi}_{MH}V_j)^2}{\left[\sum_{j=1}^{k}V_j\right]^2}. \quad \textbf{(42)}$$

Confidence interval construction about this estimator $\hat{\psi}_{MH}$ is usually based on the distribution of its logarithm (Donner and Klar, 2000), since the sampling distribution of this estimator tends to be severely non-normal. The variance of $\ell n(\hat{\psi}_{MH})$ is estimated by:

$$\hat{V}(\ell n[\hat{\psi}_{MH}]) = \frac{\hat{V}(\hat{\psi}_{MH})}{[\hat{\psi}_{MH}]^2}. \quad \textbf{(43)}$$

Therefore approximate two-sided $(1-\alpha)100$ per cent confidence limits about ψ, which need large number of strata (at least 15), are given, after approximate transformation, by:

$$\exp\{\ell n(\hat{\psi}_{MH}) \pm Z_{\alpha/2}[\hat{V}(\ell n[\hat{\psi}_{MH}])^{1/2}]\}. \quad \textbf{(44)}$$

Chapter 4

Analysis of Data and Interpretation of Results

The 'DOTSPack' data set and "Stepping Stones" data sets described in Chapter 1 are used to demonstrate the application of the aggregated analysis, disaggregated and adjusted chi-square approaches, under sections 4.1, 4.2 and 4.3 respectively.

4.1 Aggregated Analysis

In order to perform the cluster-level analysis, summary statistics for each cluster were created, for both DOTSPack and "Stepping Stones" data sets.

4.1.1 Analysis of 'DOTSPack'

The cross-tabulation of group and cure shows a marginal cure-rate difference of about 8% in favour of DOTSPack (Table 15, Appendix). That is, overall DOTSPacks arm seems to be having a higher cure rate by about 8% as illustrated in Figure 7.

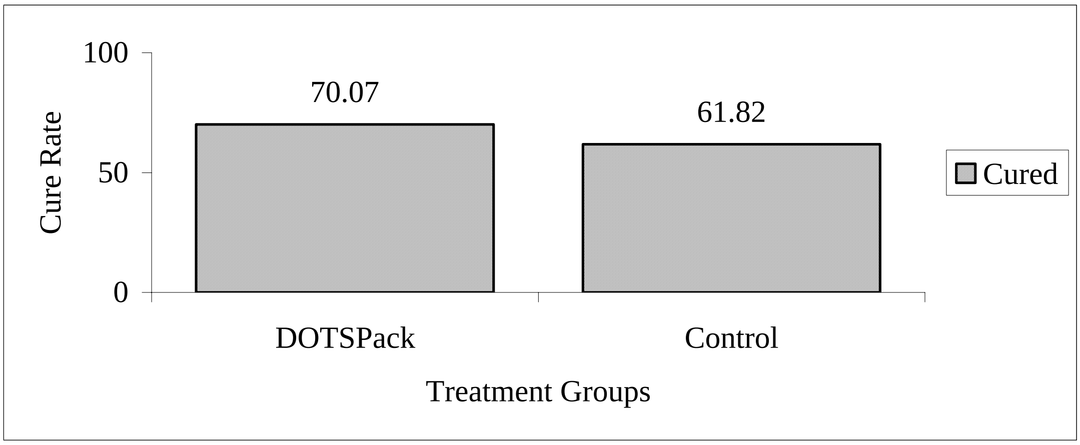

Figure 7: Representation of the overall cure rates

Table 13 (Appendix) presents the DOTSPack's collapsed data set. The results from a weighted one-way **an**alysis **o**f **va**riance (ANOVA), with the number of subjects within a clinic being the weights, are presented by Table 5 , from which it can be concluded that the null hypothesis of no cure rate difference between the two arms is not rejected ($p-value = 0.1806, F = 1.86$). Thus, there is not sufficient evidence to conclude that the DOTSPack has a higher cure rate than the control.

Also, Table 5 shows that it can be predicted that the DOTSPack's cure rate would increase by about 0.083 more than that of the controls.

Table 5: DOTSPack Predicted coefficients - oneway ANOVA

```
Source        |  SS              df        MS         Number of obs =      39
--------------+------------------------------         F(  1,     37) =    1.86
       Model |  .065326378      1   .065326378        Prob > F      =  0.1806
    Residual |  1.29787002     37   .035077568        R-squared     =  0.0479
--------------+------------------------------         Adj R-squared =  0.0222
       Total |   1.3631964     38   .035873589        Root MSE      =  .18729

--------------------------------------------------------------------------
  cured          Coef.    Std. Err.       t     P>|t|      [95% Conf. Interval]
--------------------------------------------------------------------------
cons          .6181818    .0453927     13.62    0.000      .5262074     .7101562
group
     1      .0825224    .0604703      1.36    0.181     -.0400021      .205047
     2      (dropped)
      --------------------------------------------------------------
```

4.1.2 'Stepping Stones'

Table 6 shows that controls seem to be having higher prevalence of HIV and 'correct condom use' than the stones, while stones seem to be having higher prevalence in 'any physical abuse in the past year' than the controls. There are marginal differences of 3.01% and 3.28% for 'HIV' and 'correct condom use', respectively, in favour of controls, while a marginal difference of about 2.7% for 'any physical abuse' is in favour of stones.

Table 6: Proportions - comparisons between Stones and controls

Treat	**HIV Status**		**Any physical abuse in past year**		**Correct condom use last sex**	
	0	**1**	**0**	**1**	**0**	**1**
Control	610 ***87.02***	91 ***12.98***	501 ***72.93***	186 ***27.07***	386 ***60.41***	253 ***39.59***
Stones	623 ***90.03***	69 ***9.97***	472 ***70.24***	200 ***29.76***	407 ***63.69***	232 ***36.31***
Total	1233 88.51	160 11.49	973 71.60	386 26.40	793 62.05	485 37.95

The null hypothesis that prevalence for ‘HIV’, between the two arms is not statistically different is not rejected at 95% level of significant (*p*-value = 0.1381). So are the hypotheses that the prevalence of ‘any physical abuse’ and ‘correct condom use’ between the two arms are not statistically different are not rejected, since the estimated significance levels are 0.3689 and 0.3959, respectively. Table 21 (Appendix) shows the results from the weighted one-way analysis of variance, where the numbers of subjects within clusters were used as weights. Figure 8 represents the insignificant marginal differences between the two arms.

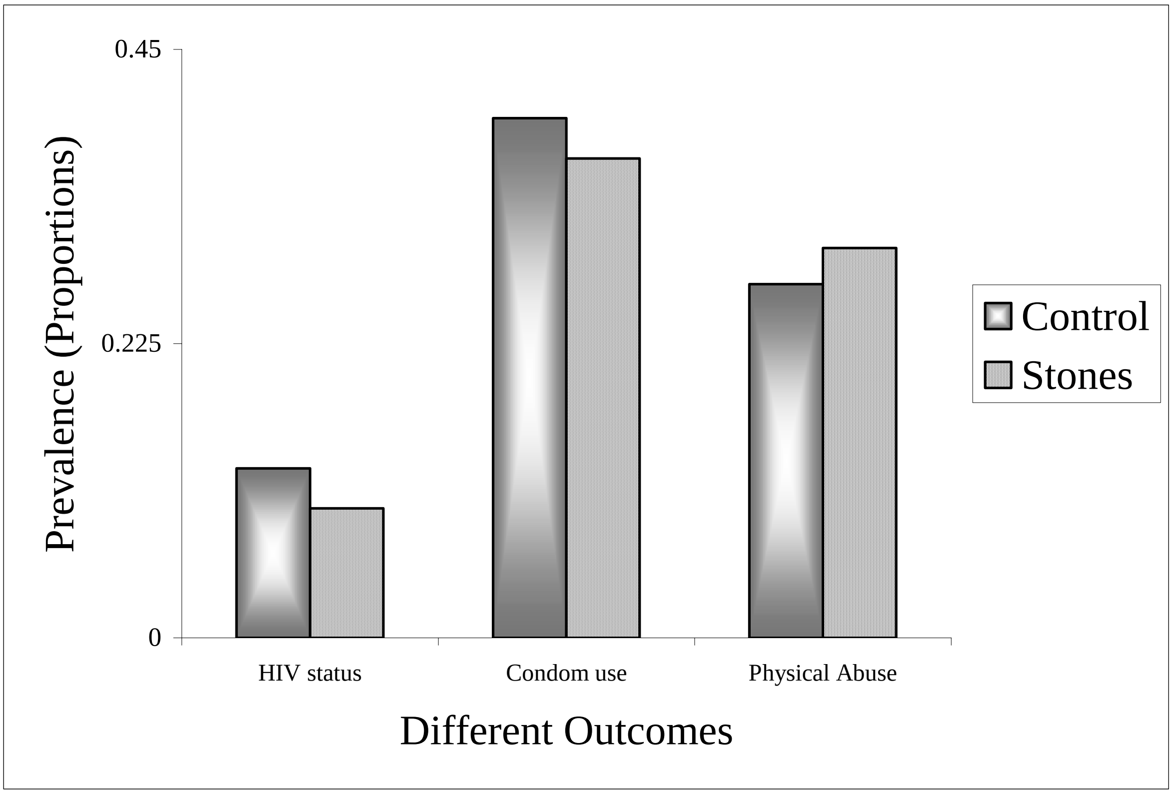

Figure 8: Weighted one-way ANOVA - marginal differences

4.2 Disaggregated Analysis

The DOTSPack data set will be used to demonstrate the application of GLMM (Generalised Linear Mixed Models) and GEE (Generalised Estimating Equations) under sections 4.2.1 and 4.2.2, respectively, while the "Stepping Stones" data set will be discussed under sections 4.2.3 and 4.2.4 to demonstrate the GLMM and GEE respectively.

4.2.1 Generalised Linear Mixed Models (GLMM)

The estimated intraclass correlation is $\hat{\rho} = 0.09436$ (Table 14, Appendix), therefore the variation between patients in different clinics exceeds variation within clinics; hence the design is characterized by 'between cluster variation'.

The fitted model shows that the cure rate scores are significantly different for the groups, control group versus DOTSPack group. DOTSPack have significantly higher cure rates than the control group. Based on the results, the odds of cure in the DOTS Pack group are 2.80 times as high as the odds of cure in the control. Adjusting for 'extra-pulmonary TB' and ' baseline smear status' does not change the general conclusion namely that the odds of cure in the DOTS Pack group are about 2.86 times as high as the odds of cure in the control group.

Table 24 (Appendix) shows the fitted models where model **(1)** is unadjusted model and in model **(2)** we adjusted for covariates 'extra-pulmonary' TB and 'baseline smear status'.

Table 7 shows that adjusting for baseline smear status and whether or not the patient had extra-pulmonary TB, controls are significantly less likely to be cured than DOTSPack (odds ratio = 0.35). That is, controls are about 0.35 less likely to be cured.

Table 7: GLMM - Odds Ratio

```
Random-effects logistic regression          Number of obs      =       504
Group variable (i): clincod                 Number of groups   =        39

Random effects u_i ~ Gaussian               Obs per group: min =         1
                                                           avg =      12.9
                                                           max =       108

                                            Wald chi2(3)       =     12.96
Log likelihood  =  -308.5191                Prob > chi2        =    0.0047

---------------------------------------------------------------------------
 cured    |         OR   Std. Err.       z     P>|z|      [95% Conf. Interval]
----------+----------------------------------------------------------------
Igroup_2 |   .3501486   .1205765    -3.05   0.002       .178293    .6876549
      ep |    2.52184   1.353292     1.72   0.085      .8809239    7.219327
 ptsmear |   .7589809    .123165    -1.70   0.089      .5522053    1.043184
----------+----------------------------------------------------------------
/lnsig2u |  -1.169563   .4666104                       -2.084102   -.2550233
----------+----------------------------------------------------------------
 sigma_u |   .5572276   .1300041                        .3527304    .8802831
     rho |   .0862419   .0367709                        .0364406    .1906379
---------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) = 18.20 Prob >= chibar2 = 0.000
```

4.2.2 Generalised Estimating Equation (GEE)

The fitted models (**a**) and (**b**), Table 18 (Appendix), show the predicted coefficients, indicating that there is a significant decrease in cure rate for controls. That is, the odds of cure in the control group are 3 times as low as the odds of cure in the DOTSPack group. Model **(a)** is an unadjusted model while in model (**b**) we have adjusted for ‘extra-pulmonary TB’ and ‘ baseline smear status’. The results from both models are almost the same. Therefore the general conclusion from the two models is the same.

It can be seen from Table 8 that adjusting for baseline smear status and whether or not the patient had extra-pulmonary TB, controls are significantly less likely to be cured than DOTSPack (odds ratio = 0.3213364). That is, being in the control group reduces the chances of being cured from TB. The odds of a control subject being cured are 0.321 times that of a DOTSPack subject.

Table 8: GEE – DOTSPack

```
GEE population-averaged model                 Number of obs        =        504
Group variable:                     clincod   Number of groups     =         39
Link:                                 logit   Obs per group: min   =          1
Family:                            binomial                  avg   =       12.9
Correlation:                   exchangeable                  max   =        108
                                                 Wald chi2(3)      =      12.88
Scale parameter:                          1      Prob > chi2       =     0.0049


---------------------------------------------------------------------------
       cured | Odds Ratio   Std. Err.    z    P>|z|     [95% Conf. Interval]
-------------+-------------------------------------------------------------
   _Igroup_2 |  .3213364 .128783     -2.83   0.005      .1464945    .704853
          ep |   2.50432 1.167224     1.97   0.049     1.004519    6.243406
      ptsmear| .7536718  .1074786    -1.98   0.047      .5698956    .996711
```

4.2.3 Generalised Linear Mixed Models (GLMM)

The estimated intraclass correlation coefficients when 'HIV status' 'any physical abuse' and 'correct condom use' are dependent variables, as presented by Table 26, are 0.02555, 0.02854 and 0.04113, respectively. Therefore the variations between subjects in different clusters exceed variations within clusters, respectively; hence the designs are characterised by 'between cluster variation'.

a) Unadjusted models

Table 19 (Appendix) shows the regression coefficients of the three unadjusted models, (**1**), (**2**) and (**3**) for HIV status, 'correct condom use' and ' any physical abuse', respectively. These results indicate that the HIV prevalence, 'correct condom use' prevalence and 'any physical abuse' prevalence are not significantly different for the two arms. The odds of a control subject being HIV positive, and 'using condom correctly' are 0.74 and 0.86 times as low as the odds in the treatment group, respectively. Otherwise, the odds of 'physical abuse' in the control group are 1.14 times as high as the odds of 'physical abuse' in the treatment group. Table 9 below presents this information.

Table 9: "Stepping Stones" Unadjusted models

```
Random-effects logistic regression           Number of obs      =      1393
Group variable (i): clustno                  Number of groups   =        69

Random effects u_i ~ Gaussian                Obs per group: min =        15
                                                            avg =      20.2
                                                            max =        30

                                             Wald chi2(1)       =      2.25
Log likelihood  = -492.94372                 Prob > chi2        =    0.1333

------------------------------------------------------------------------------
 hivstat |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
Itreat_2 |   .7363036     .15014    -1.50   0.133     .4937282    1.098059
---------+--------------------------------------------------------------------
/lnsig2u |  -1.609904   .6188716                      -2.82287   -.3969377
---------+--------------------------------------------------------------------
 sigma_u |   .4471095   .1383517                      .2437932    .8199853
     rho |   .0572836   .0334204                      .0177455    .1696957
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) = 4.37 Prob >= chibar2 = 0.018
```

```
Random-effects logistic regression           Number of obs      =      1278
Group variable (i): clustno                  Number of groups   =        69

Random effects u_i ~ Gaussian                Obs per group: min =        14
                                                            avg =      18.5
                                                            max =        29

                                             Wald chi2(1)       =      0.91
Log likelihood  = -841.39022                 Prob > chi2        =    0.3402

------------------------------------------------------------------------------
  ccondo |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
Itreat_2 |   .8603375   .1356865    -0.95   0.340     .6315725    1.171965
---------+--------------------------------------------------------------------
/lnsig2u |  -1.678485   .4214453                     -2.504503   -.8524673
---------+--------------------------------------------------------------------
 sigma_u |   .4320377   .0910401                      .2858605    .6529638
     rho |   .0536906   .0214127                      .0242367    .1147296
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) = 12.47 Prob >= chibar2 = 0.000
```

```
Random-effects logistic regression               Number of obs      =      1359
Group variable (i): clustno                      Number of groups   =        69

Random effects u_i ~ Gaussian                    Obs per group: min =        16
                                                                avg =      19.7
                                                                max =        29

                                                 Wald chi2(1)       =      0.79
Log likelihood  = -806.94268                     Prob > chi2        =    0.3745

------------------------------------------------------------------------------
      anypa1 |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Itreat_2 |   1.144911   .1744513     0.89   0.374     .8493236     1.54337
-------------+----------------------------------------------------------------
    /lnsig2u |  -1.959451   .5185537                     -2.975797   -.9431042
-------------+----------------------------------------------------------------
     sigma_u |   .3754142   .0973362                      .2258467     .624033
         rho |   .0410795   .0204269                      .0152675    .1058404
------------------------------------------------------------------------------
```

Likelihood-ratio test of rho=0: chibar2(01) = 6.80 Prob >= chibar2 = 0.005

b) Adjusted models

Adjusting for age at first sexual violence by partner, sexual abuse last year score, HIV knowledge score, attitude to condoms, age at first sex, number of sex partners last year and any drug use gives the results that are not statistically different between the two arms. That is, “Stepping Stones” and controls are not statistically different in so far as the HIV prevalence is concerned. Same conclusions were reached for correct condom use, and for any physical abuse in the past year.

From the results represented in Table 23 (Appendix), which represent the regression coefficients, prevalence of the correct condom use, HIV and any physical abuse are not significantly different for the two arms. Table 10 shows that the odds of HIVstat in the control group are 0.210 times as low as the odds of HIV in the treatment group. Otherwise, the odds of ‘correct condom use’ and ‘any abuse in the past year’ in the control group were 1.3991 and 1.4955 times as high as the odds in the treatment group, respectively.

Table 10: GLMM Adjusted models ("Stepping Stones")

```
Random-effects logistic regression              Number of obs      =       104
Group variable (i): clustno                     Number of groups   =        56

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       1.9
                                                               max =         5

                                                Wald chi2(8)       =      9.31
Log likelihood  = -57.858734                    Prob > chi2        =    0.3165

------------------------------------------------------------------------------
   ccondo |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Itreat_2 |   1.399063   .7146974     0.66   0.511     .5140549    3.807721
    sayr1 |    .777391   .1040359    -1.88   0.060     .5980339    1.010539
   psvage |   .9812486   .1022015    -0.18   0.856     .8000599    1.203471
  agefsex |   .8914335   .1173477    -0.87   0.383     .6887115    1.153827
 numpart1 |   1.207405    .216184     1.05   0.293      .850054     .71498
Idruguse_1|   2.091632    2.05823     0.75   0.453     .3040022   14.39109
 condoatt |   1.092221   .0595189     1.62   0.105     .9815796    1.215333
  hivknow |   1.043829   .0559574     0.80   0.424     .9397197    1.159473
----------+-------------------------------------------------------------------
 /lnsig2u |        -14   1433.236                     -2823.092    2795.092
----------+-------------------------------------------------------------------
  sigma_u |   .0009119   .6534712                             0           .
      rho |   2.53e-07   .0003623                             0           .
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.00 Prob >= chibar2 = 1.000
```

```
Random-effects logistic regression              Number of obs      =       104
Group variable (i): clustno                     Number of groups   =        56
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       1.9
                                                               max =         5
                                                Wald chi2(8)       =      6.71
Log likelihood  =   -31.7819                    Prob > chi2        =    0.5681

------------------------------------------------------------------------------
  hivstat |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Itreat_2 |   .2102329   .1710376    -1.92   0.055     .0426771    1.035634
    sayr1 |   .8498711   .1595949    -0.87   0.386     .5881774    1.227999
   psvage |   1.632474   .4035535     1.98   0.047     1.005602    2.650125
  agefsex |   .6843273   .1505404    -1.72   0.085     .4446455    1.053207
 numpart1 |   1.260321   .3487064     0.84   0.403     .7327739    2.167665
Idruguse_1| 1.034651   1.512165     0.02   0.981     .0589853    18.14863
 condoatt |   1.017948   .0885766     0.20   0.838     .8583379    1.207238
 hivknow  |   .9452356   .0819582    -0.65   0.516     .7975083    1.120327
--------- +-------------------------------------------------------------------
/lnsig2u  |   -.623591   3.713274                     -7.901475    6.654293
--------- +-------------------------------------------------------------------
 sigma_u  |   .7321312   1.359302                      .0192405    27.85874
     rho  |   .1401026   .4473524                      .0001125     .995779
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.09 Prob >= chibar2 = 0.380
```

```
Random-effects logistic regression              Number of obs      =       104
Group variable (i): clustno                     Number of groups   =        56
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       1.9
                                                               max =         5
                                                Wald chi2(8)       =     14.32
Log likelihood  = -57.807052                    Prob > chi2        =    0.0737

------------------------------------------------------------------------------
   anypa1 |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Itreat_2 |   1.495545   .7626257     0.79   0.430     .5504845    4.063068
    sayr1 |   1.205637   .1445875     1.56   0.119     .9530922    1.525098
   psvage |   .8338603   .0960892    -1.58   0.115     .6652821    1.045155
  agefsex |    .956932    .121834    -0.35   0.730     .7456043    1.228157
 numpart1 |   1.467003   .2793822     2.01   0.044     1.010008    2.130774
 Idruguse1|   1.95e+15   3.44e+22     0.00   1.000            0           .
 condoatt |   .9034788   .0494118    -1.86   0.063     .8116434    1.005705
  hivknow |   1.101218   .0603525     1.76   0.079     .9890604    1.226093
----------+-------------------------------------------------------------------
 /lnsig2u |        -14   629.8715                     -1248.526    1220.526
----------+-------------------------------------------------------------------
  sigma_u |   .0009119   .2871842                      7.7e-272    1.1e+265
      rho |   2.53e-07   .0001592                             0           .
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.00 Prob >= chibar2 = 1.000
```

4.2.4 Generalised Estimating Equation (GEE)

a) Unadjusted models

The results of unadjusted population averaged models; show that the groups are statistically not different when compared over correct condom use. Same conclusions were reached when the two groups were compared over HIV status, and over any physical abuse in past year.

The results show that controls had higher ‘physical abuse’ scores than Stones. That is, the odds of ‘physical abuse’ in the control group are 1.143 times as high as the odds of ‘physical abuse’ in the treatment group. Otherwise, the odds of ‘HIV positive’ and ‘correct condom use’ in the control group are 0.741 and 0.869 times as low as the odds in the treatment group, respectively.

Table 11: Unadjusted models - GEE (odds ratios)

```
GEE population-averaged model                   Number of obs      =      1278
Group variable:                    clustno      Number of groups   =        69
Link:                                logit      Obs per group: min =        14
Family:                           binomial                     avg =      18.5
Correlation:                  exchangeable                     max =        29
                                                Wald chi2(1)       =      0.88
Scale parameter:                         1      Prob > chi2        =    0.3495

------------------------------------------------------------------------------
   ccondo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |    .869132   .1303107    -0.94   0.350     .6478334    1.166026
------------------------------------------------------------------------------

GEE population-averaged model                   Number of obs      =      1393
Group variable:                    clustno      Number of groups   =        69
Link:                                logit      Obs per group: min =        15
Family:                           binomial                     avg =      20.2
Correlation:                  exchangeable                     max =        30
                                                Wald chi2(1)       =      2.25
Scale parameter:                         1      Prob > chi2        =    0.1337

------------------------------------------------------------------------------
  hivstat | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Itreat_2 |   .7410928   .1480721    -1.50   0.134      .500957    1.096339
------------------------------------------------------------------------------

GEE population-averaged model                   Number of obs      =      1359
Group variable:                    clustno      Number of groups   =        69
Link:                                logit      Obs per group: min =        16
Family:                           binomial                     avg =      19.7
Correlation:                  exchangeable                     max =        29
                                                Wald chi2(1)       =      0.84
Scale parameter:                         1      Prob > chi2        =    0.3588

------------------------------------------------------------------------------
   anypa1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Itreat_2 |   1.143474   .1670742     0.92   0.359     .8587294    1.522637
------------------------------------------------------------------------------
```

b) Adjusted models

The population-averaged models were fitted, adjusting for age at first sexual violence by partner, sexual abuse last year score, HIV knowledge, attitude to condoms, age at first sex, number of sex partners last year and any drug use.

The results (Table 22, Appendix) show that HIV scores are significantly different for groups. Stones have significantly higher HIV scores than controls. Based on the results, the odds of HIV in the control group are 0.2249 times as low as the odds of HIV in the treatment group.

Table 22 (Appendix) shows that controls had higher 'physical abuse' and 'correct condom use' scores than stones. Precisely, the odds of 'physical abuse' and 'correct condom use' in the control group are 1.002 and 1.3477 times as high as the odds in the treatment group, respectively. Table 12 presents this information.

```
GEE population-averaged model                   Number of obs      =       104
Group variable:                    clustno      Number of groups   =        56
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       1.9
Correlation:                  exchangeable                     max =         5
                                                Wald chi2(8)       =     22.76
Scale parameter:                         1      Prob > chi2        =    0.0037

------------------------------------------------------------------------------
   anypa1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   1.002486   .3876165     0.01   0.995     .4698474    2.138947
    sayr1 |   1.228386   .1308605     1.93   0.053     .9969097    1.513609
   psvage |   .7914809   .0858516    -2.16   0.031     .6398983    .9789712
  agefsex |   1.009771    .117175     0.08   0.933     .8043567    1.267645
 numpart1 |   1.363279   .2406954     1.76   0.079     .9644945    1.926947
Idruguse_ |   283.9244   1788.286     0.90   0.370     .0012358    6.52e+07
   hivknow |  1.126254   .0541415     2.47   0.013     1.024984    1.237529
  condoatt |  .9030532   .0468523    -1.97   0.049      .815739    .9997132
```

```
GEE population-averaged model                   Number of obs      =       104
Group variable:                    clustno      Number of groups   =        56
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       1.9
Correlation:                  exchangeable                     max =         5
                                                Wald chi2(8)       =     13.71
Scale parameter:                         1      Prob > chi2        =    0.0896

------------------------------------------------------------------------------
  hivstat | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   .2248803   .1537819    -2.18   0.029     .0588662    .8590854
    sayr1 |   .8612485   .1415278    -0.91   0.363     .6240964    1.188517
   psvage |   1.545343   .2444052     2.75   0.006     1.133449    2.106917
  agefsex |   .7089511   .1250332    -1.95   0.051     .5017591    1.001699
 numpart1 |   1.220508   .2913474     0.83   0.404     .7644519    1.948636
Idruguse_1|   .9163262   1.238551    -0.06   0.948     .0647925    12.95911
  hivknow |    .941682   .0768928    -0.74   0.462     .8024162    1.105119
 condoatt |   1.018613    .084042     0.22   0.823     .8665225    1.197399
```

```
GEE population-averaged model                   Number of obs      =       104
Group variable:                    clustno      Number of groups   =        56
Link:                                logit      Obs per group: min =         1
Family:                           binomial                     avg =       1.9
Correlation:                  exchangeable                     max =         5
                                                Wald chi2(8)       =      8.39
Scale parameter:                         1      Prob > chi2        =    0.3966

------------------------------------------------------------------------------
   ccondo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   1.347752   .6534168     0.62   0.538     .5211113    3.485694
    sayr1 |    .798595   .1031322    -1.74   0.082     .6200128    1.028614
   psvage |   .9874288   .1014967    -0.12   0.902     .8072569    1.207813
  agefsex |   .8839929   .1148334    -0.95   0.343     .6852912    1.140309
 numpart1 |   1.178292   .2101725     0.92   0.358     .8306604    1.671406
Idruguse_1|   1.906045    1.86721     0.66   0.510     .2794312    13.00144
  hivknow |   1.038294   .0552497     0.71   0.480     .9354618     1.15243
 condoatt |   1.088182   .0592725     1.55   0.121     .9779964    1.210782
------------------------------------------------------------------------------
```

Table 12: Adjusted models - GEE (Odds ratios)

4.3 Adjusted chi-square approach

The DOTSPack data used will be described in section 4.3.1, while the "Stepping Stone" data will be described in section 4.3.2. The C_i 's are defined to be clustering correction factors and are given by **(32),** while $\hat{P}_i$ is the event rate as computed over all clusters in ith treatment group.

4.3.1 DOTSPack

The estimated intraclass correlation coefficient was found to be 0.09436, and $C_1 = 3.174267$ and $C_2 = 6.354$ for DOTSPacks and controls respectively. Therefore, the adjusted chi-square is:

$$\chi_A{}^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i(1-\hat{P})\hat{P}} = 0.83$$

where,

the total number of individuals in controls and DOTSpack are $M_1 = 220$ and $M_2 = 284$, respectively; and $\hat{P}_1$ and $\hat{P}_2$ were found to be 0.70007042 and 0.66181818, respectively. The overall event rate, $\hat{P}$, was 0.6646826. The null hypothesis that the proportions of subjects who are HIV positive are the same in the two arms is not rejected, for the 5% critical value is 3,85 (see Table 25).

4.3.2 "Stepping Stones"

The total number of individuals in the two arms, controls and "Stepping Stones" were 711 and 701, respectively.

i. HIV prevalence

The overall event rate was $\hat{P} = 0.1127146$, while clustering correction factors were $C_1 = 1.497488$ and $C_2 = 1.506626$ for controls and "Stepping Stones", respectively. Therefore, the adjusted chi-square is given by:

$$\chi_A{}^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i(1-\hat{P})\hat{P}} = 0.8173015,$$

where

$\hat{P}_1 = 0.129587$ and $\hat{P}_2 = 0.0988172$ were the event rates as computed in the two groups, controls and "Stepping Stones", respectively. The null hypothesis that the proportions of HIV between the two arms do not statistically differ is not rejected since the 5% critical value is 3.84 (using Table 25).

ii. Correct condom use

The clustering correction factors for controls and "Stepping Stones" were 1.800849 and 1.81559, respectively. The overall event rate was found to be $\hat{P} = 0.3831441$. Thus the adjusted chi-square is given by:

$$\chi_A{}^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i(1-\hat{P})\hat{P}} = 0.8173015.$$

The 5% critical value is 3.84 (Table 25), and this implies that the null hypothesis of no differences in proportions of the correct condom use between the two arms cannot be rejected.

iii. Any physical abuse in the past year

Adjusted chi-square was found be:

$$\chi_A{}^2 = \sum_{i=1}^{2} \frac{M_i\left(\hat{P}_i - \hat{P}\right)^2}{C_i(1-\hat{P})\hat{P}} = 0.885$$

since, the clustering correction factors were found to be 1.555707 and 1.565914 for controls and "Stepping Stones", respectively, and the overall event rate was $\hat{P} = 0.2834883$. Therefore, we could not reject the hypothesis that the proportions of subjects who had experienced physical abuse in the two arms were the same, because the 5% critical value is 3.84 (Table 25).

The aggregated and adjusted chi-square approaches are applied after data resolution has been performed. Adjustment for covariates is not permissible in the application of these approaches. The application of these approaches to our data sets, DOTSPAcks and "Stepping Stones", could not reject the hypotheses of no differences between the respective two arms.

The disaggregated approaches, GLMM and/or GEE, have the advantage that they can adjust for covariates. The GLMM and GEE approaches (adjusted models) detected the significant difference in cure-rates between the DOTSPack and the control group (p-value = 0.002; OR = 0.3501486) and (p-value = 0.005, OR = 0.3213364) respectively. Also the GEE approach detected a significant difference in HIV prevalence (p-value = 0.029; OR = 0.2248803) between the "Stepping stones" and the control group. Otherwise, in all other fitted models the GLMM and the GEE approaches could not detect any significant differences.

Chapter 5

Conclusions

The aim of this study/thesis was to explore the approaches used in the analysis of cluster randomisation trials with discrete (especially binary) outcomes. In addition, the methods of analysing data with continuous outcomes were discussed.

In cluster randomised trials the responses of individuals within clusters tend to be correlated; hence the information provided by a given sample size in a CRT is generally less than in an individually randomised trials. And this has to be taken into account in determining sample size. That is, trials that ignore cluster randomization when calculating sample size result in studies that are underpowered. This may (in part) be explained by the lack of published intra-cluster correlation coefficients.

The unit of analysis has an impact on the analysis to be carried out; since it (the unit of analysis) is the basic entity about which generalisations are about to be made based on an analysis, and for which data have been collected. Questions concerning the appropriate unit of analysis are (more) challenging when the primary target of inference is at the level of individual subjects, with the choice of randomisation unit then a matter of convenience or other practical considerations.

Standard statistical techniques are inappropriate, unless an aggregated analysis is performed at the cluster level, as they require data to be independent. Although data resolution can be done so as to apply an aggregated analysis at the cluster level, this approach is inefficient, more so that it does not allow variation at the individual level to be explored. Otherwise, the other approach could be to adjust standard methods, such as the chi-squared test, to allow for the effect of clustering.

The modeling approaches (mostly) used for regression analysis outcome binary outcome data in cluster randomisation trials are GEE and GLMM, which may be characterised as ‘population-averaged’ in that it measures the expected (marginal) change in a response as the value of the covariate increases by a unit. Both approaches estimate the same population parameter when the outcome is normally distributed, with the equivalence disappearing in the case of a binary outcome variable.

The advantage of the disaggregated models (GLMM and/or GEE) over the aggregated approach and the adjusted chi-square approach is that they are able to adjust for individual level covariates, and thus gain power.

The disaggregated approach has been shown to be more advantageous than the aggregated and adjusted approaches in that:

a. For the ‘DOTSPack’ data set, the GEE and GLMM detected a significant difference between the two groups, while the aggregate and adjusted chi-square approaches failed to detect the significant difference.
b. The GEE approach, adjusted model, detected the significant difference between the two treatment groups, Stepping Stones and control groups, when the outcome variable was HIV status. The GLMM (both adjusted and unadjusted models) and adjusted chi-square approach were unable to detect this significant difference. Otherwise, all of these approaches could not detect the significant differences between the two groups for the dependent variables ‘any physical abuse’ and ‘correct condom use’

Thus, the disaggregated approach (es) has/have shown to be ‘superior’ than the aggregated and the adjusted chi-square approaches.

The challenges when dealing with cluster randomisation trials seem to be:

i. The assessment of intraclass correlation coefficient, which has a critical role to play in the estimation of sample size; and
ii. Difficulties in assessing the value of intra-cluster correlation coefficient arise because these values often are not presented in reports, and/or even reported such estimates are often based on a small number of clusters and are design dependent.

Appendix

Table 13: DOTSPack - Collapsed data set

	group	clinic	cure rate	dummy##
1.	dotspack	ALLEMANSDRIFT C CHC	.857143	7
2.	dotspack	BEATY STREET FIXED C	.692308	13
3.	dotspack	BURGERSFORT CLINIC	.604167	48
4.	dotspack	EKANGALA COM HEALTH	.791667	24
5.	dotspack	EMGWENYA CLINIC	.8	5
6.	dotspack	EMTONYENI CLINIC	.666667	3
7.	dotspack	GROBLERSDAL CLINIC	.5	4
8.	dotspack	GROBLERSDAL MOBILE	1	3
9.	dotspack	KAMEELRIVIER A CLINI	.5	2
10.	dotspack	KLIPPLAATDRIFT CLINI	1	3
11.	dotspack	KWAGGA A	1	1
12.	dotspack	KWAGGA CHC	.692308	26
13.	dotspack	MASHISHING CLINIC	.555556	27
14.	dotspack	MHLUZI MUNICIPAL CLI	1	8
15.	dotspack	MIDDELBURG CENTRE	.823529	17
16.	dotspack	MOLOTO CLINIC	.9	10
17.	dotspack	PHOLA CHC	.785714	14
18.	dotspack	SEABE CLINIC	.857143	7
19.	dotspack	SIMUNYE MUNICIPAL CL	.5	2
20.	dotspack	SIYABUSWA COMM HEALT	.628571	35
21.	dotspack	SIYATHUTHUKA CLINIC	.6	10
22.	dotspack	WITBANK CORRECTIONAL	.666667	15
23.	control	GATEWAY CLINIC	0	2
24.	control	HENDRINA MUNICIPAL C	.571429	7
25.	control	KAMEELPOORTNEK CLINI	.3	10
26.	control	KWAZAMOKUHLE MUN CLI	.4375	16
27.	control	LEEUFONTEIN CLINIC	0	3
28.	control	LYNNVILLE MUNICIPAL	.787037	108
29.	control	MARBLE HALL MUNICIPA	.666667	3
30.	control	MHLUZI MUNICIPAL MOB	.7	10
31.	control	NOKANENG CLINIC	.5	8
32.	control	PIETERSKRAAL CLINIC	1	5
33.	control	TROYA CLINIC	0	2
34.	control	TWEEFONTEIN A CLINIC	.375	8
35.	control	VALSCHFONTEIN CLINIC	.166667	6
36.	control	VLAKLAAGTE COM HEALT	.428571	14
37.	control	WATERVAL BOVEN MOBIL	1	3
38.	control	WELTEVREDE CLINIC	.142857	7
39.	control	ZAAIPLAATS CLINIC	.625	8

dummy = number of subjects

Table 14: Intraclass correlation determination

```
                     One-way Analysis of Variance for cured:

                                                Number of obs =        504
                                                    R-squared =     0.1568

    Source                SS          df       MS             F     Prob > F
-------------------------------------------------------------------------
Between clincod        17.616693     38     .46359718      2.28      0.0000
Within clincod         94.714656    465     .20368743
-------------------------------------------------------------------------
Total                  112.33135    503     .22332276

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.09438      0.04426       0.00764     0.18113

         Estimated SD of clincod effect          .1457001
         Estimated SD within clincod             .4513174
         Est. reliability of a clincod mean       .560637
              (evaluated at n=12.24)
```

Table 15: DOTSPack - group vs. cure

```
                          CURED
     group |          0          1 |     Total
-----------+----------------------+----------
  DOTSPack |         85        199 |       284
           |      29.93      70.07 |    100.00
-----------+----------------------+----------
   control |         84        136 |       220
           |      38.18      61.82 |    100.00
-----------+----------------------+----------
     Total |        169        335 |       504
           |      33.53      66.47 |    100.00
```

Table 16: GLMM - Comparison model

```
Random-effects logistic regression            Number of obs      =       504
Group variable (i): clincod                   Number of groups   =        39

Random effects u_i ~ Gaussian                 Obs per group: min =         1
                                                             avg =      12.9
                                                             max =       108

                                              Wald chi2(3)       =     12.96
Log likelihood  =  -308.5191                  Prob > chi2        =    0.0047

------------------------------------------------------------------------------
 cured     |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
_Igroup_2 |  -1.049398   .3443581    -3.05   0.002    -1.724327   -.3744681
       ep |   .9249889   .5366287     1.72   0.085    -.126784    1.976762
  ptsmear |  -.2757787   .1622768    -1.70   0.089    -.5938353    .042278
    _cons |   1.353152   .3040188     4.45   0.000     .7572862   1.949018
-----------+------------------------------------------------------------------
 /lnsig2u |  -1.169563   .4666104                     -2.084102   -.2550233
-----------+------------------------------------------------------------------
  sigma_u |   .5572276   .1300041                      .3527304    .8802831
      rho |   .0862419   .0367709                      .0364406    .1906379
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) = 18.20 Prob >= chibar2 = 0.000
```

Table 17: GEE Population-Averaged Model

```
. xi: xtlogit  cured i.group ep ptsmear, i(clincod) pa


GEE population-averaged model                   Number of obs      =       504
Group variable:                      clincod    Number of groups   =        39
Link:                                  logit    Obs per group: min =         1
Family:                             binomial                   avg =      12.9
Correlation:                    exchangeable                   max =       108
                                                Wald chi2(3)       =     12.88
Scale parameter:                           1    Prob > chi2        =    0.0049

------------------------------------------------------------------------------
    cured |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
_Igroup_2 |  -1.135267   .4007731    -2.83   0.005    -1.920767   -.3497657
       ep |   .9180174   .4660843     1.97   0.049      .004509    1.831526
  ptsmear |  -.2827983   .1426066    -1.98   0.047    -.5623021   -.0032946
    _cons |   1.386591   .3451122     4.02   0.000     .7101832    2.062998
------------------------------------------------------------------------------
```

Table 18: GEE – DOTSPack

	(a)		(b)
	Undjusted Model		Covariates-Adjusted
group==2	-1.119		-1.1353
	(2.78)**		(2.83)**
		ep	0.9180
			(1.97)*
		ptsmear	-0.2828
			(1.98)*
Constant	1.036		1.3866
	(3.68)**		(4.02)**
Observations	504		504
Number of clincod	39		39

Absolute value of z-statistics in parentheses

**** significant at 5%; ** significant at 1%***

Table 19: GLMM - Unadjusted models

	(1)	(2)	(3)
	hivstat	correct condom use last sex	any physical abuse in past year
treat==2	-0.306	-0.150	0.135
	(1.50)	(0.95)	(0.89)
Constant	-1.971	-0.440	-1.021
	(13.81)**	(3.98)**	(9.39)**
Observations	1393	1278	1359
Number of clustno	69	69	69

Absolute value of z-statistics in parentheses

**** significant at 5%; ** significant at 1%***

Table 20: GLMM (Adjusted models - Odds Ratios)

```
Random-effects logistic regression               Number of obs      =       104
Group variable (i): clustno                      Number of groups   =        56
Random effects u_i ~ Gaussian                    Obs per group: min =         1
                                                                avg =       1.9
                                                                max =         5
                                                 Wald chi2(8)       =     14.32
Log likelihood  = -57.807052                     Prob > chi2        =    0.0737

------------------------------------------------------------------------------
   anypa1 |         OR   Std. Err.      z    P>|z|      [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   1.495545   .7626257     0.79   0.430      .5504845    4.063068
    sayr1 |   1.205637   .1445875     1.56   0.119      .9530922    1.525098
   psvage |   .8338603   .0960892    -1.58   0.115      .6652821    1.045155
  agefsex |    .956932    .121834    -0.35   0.730      .7456043    1.228157
 numpart1 |   1.467003   .2793822     2.01   0.044      1.010008    2.130774
 _Idrugus |   1.95e+15   3.44e+22     0.00   1.000             0           .
 condoatt |   .9034788   .0494118    -1.86   0.063      .8116434    1.005705
  hivknow |   1.101218   .0603525     1.76   0.079      .9890604    1.226093
----------+-------------------------------------------------------------------
 /lnsig2u |        -14   629.8715                       -1248.526    1220.526
----------+-------------------------------------------------------------------
  sigma_u |   .0009119   .2871842                       7.7e-272    1.1e+265
      rho |   2.53e-07   .0001592                              0           .
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.00 Prob >= chibar2 = 1.000

Random-effects logistic regression               Number of obs      =       104
Group variable (i): clustno                      Number of groups   =        56
Random effects u_i ~ Gaussian                    Obs per group: min =         1
                                                                avg =       1.9
                                                                max =         5
                                                 Wald chi2(8)       =      6.71
Log likelihood  =   -31.7819                     Prob > chi2        =    0.5681

------------------------------------------------------------------------------
  hivstat |         OR   Std. Err.      z    P>|z|      [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   .2102329   .1710376    -1.92   0.055      .0426771    1.035634
    sayr1 |   .8498711   .1595949    -0.87   0.386      .5881774    1.227999
   psvage |   1.632474   .4035535     1.98   0.047      1.005602    2.650125
  agefsex |   .6843273   .1505404    -1.72   0.085      .4446455    1.053207
 numpart1 |   1.260321   .3487064     0.84   0.403      .7327739    2.167665
Idruguse_ |   1.034651   1.512165     0.02   0.981      .0589853    18.14863
 condoatt |   1.017948   .0885766     0.20   0.838      .8583379    1.207238
  hivknow |   .9452356   .0819582    -0.65   0.516      .7975083    1.120327
----------+-------------------------------------------------------------------
 /lnsig2u |   -.623591   3.713274                       -7.901475    6.654293
----------+-------------------------------------------------------------------
  sigma_u |   .7321312   1.359302                        .0192405    27.85874
      rho |   .1401026   .4473524                        .0001125     .995779
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.09 Prob >= chibar2 = 0.380

Random-effects logistic regression               Number of obs      =       104
Group variable (i): clustno                      Number of groups   =        56
Random effects u_i ~ Gaussian                    Obs per group: min =         1
                                                                avg =       1.9
                                                                max =         5
                                                 Wald chi2(8)       =      9.31
Log likelihood  = -57.858734                     Prob > chi2        =    0.3165

------------------------------------------------------------------------------
   ccondo |         OR   Std. Err.      z    P>|z|      [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Itreat_2 |   1.399063   .7146974     0.66   0.511      .5140549    3.807721
    sayr1 |    .777391   .1040359    -1.88   0.060      .5980339    1.010539
   psvage |   .9812486   .1022015    -0.18   0.856      .8000599    1.203471
  agefsex |   .8914335   .1173477    -0.87   0.383      .6887115    1.153827
 numpart1 |   1.207405    .216184     1.05   0.293       .850054    1.714981
Idruguse_ |   2.091632    2.05823     0.75   0.453      .3040022    14.39109
 condoatt |   1.092221   .0595189     1.62   0.105      .9815796    1.215333
  hivknow |   1.043829   .0559574     0.80   0.424      .9397197    1.159473
----------+-------------------------------------------------------------------
 /lnsig2u |        -14   1433.236                       -2823.092    2795.092
----------+-------------------------------------------------------------------
  sigma_u |   .0009119   .6534712                              0           .
      rho |   2.53e-07   .0003623                              0           .
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   0.00 Prob >= chibar2 = 1.000
```

Table 21: One-way ANOVA - '“Stepping Stones”'

treat	Summary of (mean) hivstat Mean	Std. Dev.	Freq.	Obs.
control	.12958701	.09040845	711	35
stones	.09895841	.07859228	700	34
Total	.1143921	.08553208	1411	69

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	.016181431	1	.016181431	2.25	0.1381
Within groups	.481288637	67	.007183412		
Total	.497470068	68	.007315736		

Bartlett's test for equal variances: chi2(1) = 0.5845 Prob>chi2 = 0.445

treat	Summary of (mean) anypal Mean	Std. Dev.	Freq.	Obs.
control	.27040595	.11575924	711	35
stones	.29813805	.13808082	700	34
Total	.2841639	.12714795	1411	69

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	.013265647	1	.013265647	0.82	0.3689
Within groups	1.08606322	67	.016209899		
Total	1.09932887	68	.016166601		

Bartlett's test for equal variances: chi2(1) = 1.1013 Prob>chi2 = 0.294

treat	Summary of (mean) ccondo Mean	Std. Dev.	Freq.	Obs.
control	.39743192	.13845939	711	35
stones	.36659224	.16071871	700	34
Total	.38213229	.14961301	1411	69

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	.016405229	1	.016405229	0.73	0.3959
Within groups	1.50571037	67	.022473289		
Total	1.5221156	68	.022384053		

Bartlett's test for equal variances: chi2(1) = 0.7988 Prob>chi2 = 0.371

Table 22: Adjusted models - Population Averaged Model

	(1) any physical abuse in past year	(2) hivstat	(3) correct condom use last sex
treat==2	**0.002**	**-1.492**	**0.298**
	(0.01)	**(2.18)***	**(0.62)**
sayr1	0.206	-0.149	-0.225
	(1.93)	(0.91)	(1.74)
psvage	-0.234	0.435	-0.013
	(2.16)*	(2.75)**	(0.12)
agefsex	0.010	-0.344	-0.123
	(0.08)	(1.95)	(0.95)
numpart1	0.310	0.199	0.164
	(1.76)	(0.83)	(0.92)
druguse==1	5.649	-0.087	0.645
	(0.90)	(0.06)	(0.66)
HIVknow	0.119	-0.060	0.038
	(2.47)*	(0.74)	(0.71)
condoatt	-0.102	0.018	0.085
	(1.97)*	(0.22)	(1.55)
Constant	1.084	-1.682	-0.832
	(0.47)	(0.49)	(0.35)
Observations	104	104	104
Number of clustno	56	56	56

Absolute value of z-statistics in parentheses
** significant at 5%; ** significant at 1%*

Table 23: GLMM - Adjusted models

	(1)	(2)	(3)
	correct condom use last sex	hivstat	any physical abuse in past year
treat==2	0.336	-1.560	0.402
	(0.66)	(1.92)	(0.79)
sayr1	-0.252	-0.163	0.187
	(1.88)	(0.87)	(1.56)
psvage	-0.019	0.490	-0.182
	(0.18)	(1.98)*	(1.58)
agefsex	-0.115	-0.379	-0.044
	(0.87)	(1.72)	(0.35)
numpart1	0.188	0.231	0.383
	(1.05)	(0.84)	(2.01)*
druguse==1	0.738	0.034	35.206
	(0.75)	(0.02)	(0.00)
condoatt	0.088	0.018	-0.102
	(1.62)	(0.20)	(1.86)
HIVknow	0.043	-0.056	0.096
	(0.80)	(0.65)	(1.76)
Constant	-0.987	-2.309	1.475
	(0.42)	(0.56)	(0.58)
Observations	104	104	104
Number of clustno	56	56	56

Absolute value of z-statistics in parentheses

**** significant at 5%; ** significant at 1%***

Table 24: GLMM – DOTSPacks

	(1) DOTSPack		(2) DOTSPack: adjusting for covariates
group==2	-1.028		-1.0494
	(3.09)**		(3.05)**
		ep	0.9250
			(1.72)
		ptsmear	-0.2758
			(1.70)
Constant	1.018		1.3532
	(4.99)**		(4.45)**
Observations	504		504
Number of clincod	39		39

Absolute value of z-statistics in parentheses

**** significant at 5%; ** significant at 1%***

Table 25: Chi-square distribution

Values of the χ^2 distribution exceeded with probability P.

d.f. \ P	0·995	0·975	0·050	0·025	0·010	0·005	0·001
1	$3{\cdot}9 \times 10^{-5}$	$9{\cdot}8 \times 10^{-4}$	3·84	5·02	6·63	7·88	10·83
2	0·010	0·051	5·99	7·38	9·21	10·60	13·81
3	0·071	0·22	7·81	9·35	11·34	12·84	16·27
4	0·21	0·48	9·49	11·14	13·28	14·86	18·47
5	0·41	0·83	11·07	12·83	15·09	16·75	20·52
6	0·68	1·24	12·59	14·45	16·81	18·55	22·46
7	0·99	1·69	14·07	16·01	18·48	20·28	24·32
8	1·34	2·18	15·51	17·53	20·09	21·96	26·13
9	1·73	2·70	16·92	19·02	21·67	23·59	27·88
10	2·16	3·25	18·31	20·48	23·21	25·19	29·59
11	2·60	3·82	19·68	21·92	24·73	26·76	31·26
12	3·07	4·40	21·03	23·34	26·22	28·30	32·91
13	3·57	5·01	22·36	24·74	27·69	29·82	34·53
14	4·07	5·63	23·68	26·12	29·14	31·32	36·12
15	4·60	6·26	25·00	27·49	30·58	32·80	37·70
16	5·14	6·91	26·30	28·85	32·00	34·27	39·25
17	5·70	7·56	27·59	30·19	33·41	35·72	40·79
18	6·26	8·23	28·87	31·53	34·81	37·16	42·31
19	6·84	8·91	30·14	32·85	36·19	38·58	43·82
20	7·43	9·59	31·41	34·17	37·57	40·00	45·32
21	8·03	10·28	32·67	35·48	38·93	41·40	46·80
22	8·64	10·98	33·92	36·78	40·29	42·80	48·27
23	9·26	11·69	35·17	38·08	41·64	44·18	49·73
24	9·89	12·40	36·42	39·36	42·98	45·56	51·18
25	10·52	13·12	37·65	40·65	44·31	46·93	52·62
26	11·16	13·84	38·89	41·92	45·64	48·29	54·05
27	11·81	14·57	40·11	43·19	46·96	49·64	55·48
28	12·46	15·31	41·34	44·46	48·28	50·99	56·89
29	13·12	16·05	42·56	45·72	49·59	52·34	58·30
30	13·79	16·79	43·77	46·98	50·89	53·67	59·70
40	20·71	24·43	55·76	59·34	63·69	66·77	73·40
50	27·99	32·36	67·50	71·42	76·16	79·49	86·66
60	35·53	40·48	79·08	83·30	88·38	91·95	99·61
70	43·28	48·76	90·53	95·02	100·43	104·22	112·32
80	51·17	57·15	101·88	106·63	112·33	116·32	124·84
90	59·20	65·65	113·15	118·14	124·12	128·30	137·21
100	67·33	74·22	124·34	129·56	135·81	140·17	149·44

For degrees of freedom $f > 100$, test $\sqrt{2\chi^2_{(f)}}$ as $\mathcal{N}(\sqrt{2f-1},\ 1)$.

Table 26: "Stepping Stones" Estimated ICC

```
               One-way Analysis of Variance for hivstat:
                                             Number of obs =      1413
                                                 R-squared =    0.0728

    Source                SS          df       MS            F     Prob > F
------------------------------------------------------------------------
Between clustno        10.335496     69     .1497898      1.53     0.0041
Within clustno         131.54702   1343    .09795013
------------------------------------------------------------------------
Total                  141.88252   1412    .10048337

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.02555     0.01257      0.00092      0.05019

         Estimated SD of clustno effect          .050681
         Estimated SD within clustno            .3129699
         Est. reliability of a clustno mean      0.34608
              (evaluated at n=20.18)
      One-way Analysis of Variance for ccondo: correct condom use last sex

                                             Number of obs =      1297
                                                 R-squared =    0.0917

    Source                SS          df       MS            F     Prob > F
------------------------------------------------------------------------
Between clustno        28.036711     69    .40632915      1.79     0.0001
Within clustno         277.80909   1227    .22641327
------------------------------------------------------------------------
Total                   305.8458   1296    .23599213

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.04113     0.01561      0.01054      0.07173

         Estimated SD of clustno effect         .0985528
         Estimated SD within clustno             .475829
         Est. reliability of a clustno mean      0.44278
              (evaluated at n=18.52)

    One-way Analysis of Variance for anypa1: any physical abuse in past year

                                             Number of obs =      1379
                                                 R-squared =    0.0768

    Source                SS          df       MS            F     Prob > F
------------------------------------------------------------------------
Between clustno        21.519394     69    .31187527      1.58     0.0022
Within clustno         258.61694   1309    .19756832
------------------------------------------------------------------------
Total                  280.13633   1378    .20329197

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.02854     0.01322      0.00262      0.05445

         Estimated SD of clustno effect         .0761804
         Estimated SD within clustno            .4444866
         Est. reliability of a clustno mean        0.366
              (evaluated at n=19.70)
```

Reference

Agresti, A. (1996). *An introduction to categorical Data Analysis.* John Wiley and Sons, Inc.

Altman, D. G. (1991). *Practical statistics for medical research.* Chapman and Hall. New York.

Altman, D. G, Bland, J. M. (1997) *Statistics notes: Unit of analysis.* BMJ **314**: 1874.

Bennett, S., Parpia, T., Hayes, R. Cousens, S. (2002) *Methods for the analysis of incidence rates in cluster randomised trials*. International Journal of Epidemiology **31**: 839 – 846.

Burton, P., Gurin, L., Sly, P. (1998)*Tutorial in biostatistics Extending the simple linear regression model to account for correlated responses: An introduction to generalised estimating equations and multi-level mixed modeling.* Statistics in Medicine **17**: 1261 – 1291.

Campbell, M. K., Grimshaw, J. M. (1998). *Cluster randomized trials: time for improvement.* BMJ **317**: 1171 - 1172

Chuang, J., Hripcsak, G., Jenders, R. A. (2000). *Considering Clustering: A methodological review of clinical Decision Support System Studies.* Proc AMIA Symp 146 –150.

Cochran, W. G. (1977). *Sampling techniques (3rd Edition).* New York: Wiley.

Cochran, W. G., Cox, G. M. (1957). *Experimental Designs (2nd Edition).* New York: Wiley.

Conover, W. J. (1980). *Practical Nonparametric Statistics 2nd edition.* John Wiley. New York

Cornfield, J. (1978). *Randomization by group: A formal analysis.* American Journal of Epidemiology **108**: 100 – 102

Diggle, P. J, Liang, K., Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford Press. New York

Donald, A, Donner, A. (1987) *Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered* Statistics in Medicine **6**: 491 - 499

Donner, A. (1985). *A regression approach to the analysis of data arising from cluster randomisation.* Int. J. Epidemiol **2**: 322 - 326

Donner A. (1992) .*Sample size requirement for stratified cluster randomisation designs.* Stat Med **11**: 743-750

Donner, A. (1998). *Some aspects of the design and analysis of cluster randomisation trials.* Appl. Statist **47**: 95 – 113

Donner, A., Birkett, N., Buck, C. (1981). *Randomisation by cluster: Sample size requirements and analysis*. American Journal of Epidemiology **114**: 906 - 914

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

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

Donner, A., Eliasziw, M., Klar, N. (1994). *A comparison of methods for Testing Homogeneity of Proportions in Teratologic studies.* Statistics in medicine **13**: 1253 - 1264.

Donner, A., Klar, N. (1994). *Cluster randomisation trials in epidemiology: Theory and application.* Journal of Stat Planning and inference **42**: 37 - 56

Donner, A., Klar, N. (2000): *Design and analysis of Cluster randomisation Trials in Health research. Arnold:* London.

Donner, A., Koval, J. J. (1980). *The estimation of intraclass correlation in the analysis of family data.* Biometrics **36**: 19 – 2

Diehr, P. D., Martin, D. C., 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

Everitt, B.S. (1993). *Cluster Analysis (3rd Edition)*. John Wiley. New York.

Farnum, N. R. (1994). *Modern Statistical Quality Control and Improvement.* Duxbury Press, California

Field, A. (2000). *Discovering statistics using SPSS for windows.* SAGE, London

FitzGerald, P. E. B. (2002). *Extended Generalized Estimating Equations for Binary Familial Data with Incomplete Families.* Biometrics **58**: 718 - 726

Goldstein, H. (1995). *Multilevel statistical Models (2nd Edition).* Arnold: London.

Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1998). *Multivariate data analysis (5th Edition).* Prentice Hall, New Jersey

Hauch, W. W., Gilliss, C. L., Donner, A., Gortner, S. (1991). Methodology Corner: *Randomisation by cluster.* Nursing Research **40**: 356 – 358.

Hayes, R. J., Bennett, S. (1999). *Simple sample size calculation for cluster-randomised trials.* International Journal of Epidemiology**28**: 319 – 326.

Heeren, T., D'agostino, R. (1987). *Robustness of the two independent t-test when applied to ordinal scaled data.* Statistics in Medicine **6**: 80 - 90

Hopkins, K. D. (1982). *The Unit of Analysis: Group Means versus Individual Observations.* American Educational research journal **19**: 5 – 18.

Hosmer, D. W., Lemeshow, S. (1989). *Applied logistic regression.* John Wiley and Sons. New York.

Hosmer, D. W., Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* John Wiley and Sons, Inc.

Hox, J. J. (1995) *Applied multilevel analysis.* Amsterdam: TT-Publkaties.

Hsieh, F. Y. (1988) *Sample size formulae for intervention studies with the cluster as unit of randomisation*. Stat Med **8**: 1195 – 1201.

Isaakidid, P., Loannidid, J. P. A. (2003). *Evaluation of Cluster Randomised Controlled Trials in Sub-Saharan Africa.* American Journal of Epidemiology **158**: 921 – 926.

Johnson, R. A., Wichern, D. W. (1992). *Applied multivariate statistical analysis.* Prentice-Hall, Inc.

Kirk, R. E. (1995). *Experimental design: Procedure for the behavioral sciences (3^{rd} ed).* Pacific Grove, CA: Books/Cole.

Klar, N., Donner, A. (1997). *The merits of matching in community intervention trials: A cautionary Tale.* Stat Med **16**: 1753 – 1764

Korn, E. L., Graubard, B. I. (1999). *Analysis of Health Surveys.* John Wiley and Sons, Inc.

Kuehl, R. O. (2000). *Design of experiments: Statistical Principles of Research Design and Analysis (2nd ed).* Pacific Grove, CA: Books/Cole.

Kupper, L. L., Hafner, K. B. (1989). *How appropriate are popular sample size formulas?.* The American Statistician **43**: 101 – 105

Langford, I. H., Bentham, G., McDonald, A. (1998). *Multi-level modeling of geographically aggregated health data: A case study on malignant melanoma mortality and UV exposure in the European community.* Stat Med **17**:41 – 57.

Lake, S., Kammann, E., Klar, N., Batensky, R. (2002). *Sample size re-estimation in cluster randomisation trials.* Stat Med **21**: 1337 – 1350

Larson, H. J. (1982). *Introduction to probability theory and statistical inference.* John Wiley & Sons, New York.

Levy, S. P., Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications (3rd Edition)*. New York: Wiley.

Lin, W., Raz, J., Harlow, S. D. (1997). *Linear Mixed Models with Heterogeneous Within-Cluster Variance.* Biometrics **53**: 190 - 923

Longford, N. T. (1993). *Random Coefficient Models.* New Jersey:Princeton

MacFarland, T. W. (1998). *Wilcoxon Matched-Paires Signed Ranks Test.* Unpublished.

Martin, D. C., Diehr, P., Perrin, E. D., Koepsell, T. D. (1993). *The effect of matching on the power of randomised community intervention studies.* Stat Med **12**: 329 – 338.

Mendenhall, W., Scheaffer, R. L., Wackerly, D. D. (1981). *Mathematical Statistics with Applications (2nd edition)*. Duxbury Press. Boston

Montgomery, D. C., Peck, E. A. (1982). Introduction to *Linear Regression Analysis.* New York: Wiley

MRC-UK. (2002). *Cluster randomised trials: Methodological and ethical considerations.* Medical Research Council: London

Murry, D. M. (1998). *Design and Analysis of Group-Randomized Trials.* Oxford University Press, Inc.

Murry, D. M. (2001). *Statistical Models Appropriate for Designs Often Used in Group-Randomized Trials.* Statist. Med. **20**: 1373 - 1385

Omar, R. Z., Thompson, S. G. (2000). *Analysis of cluster randomized trial with binary outcome data using a multi-level model.* Statist. Med. **19**: 2675 - 2688

Osborne, J. W. (2000). *The advantages of Hierarchical Linear Modeling. ERIC/AE Digest.* http://www.ericfacility.net/databases/ERIC_Digests/ed447198.html

Pearl, J. *Simpson's Paradox: An Anatomy.* University of California. http://singpore.cs.ucla.edu/R264.pdf

Pocock, S. J. (1983): *Clinical trials. A practical approach*. New York: John Wiley.

Powers, D. A., Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis.* Academic Press, London.

Rao, J. N. K., Scott, A. J. (1992) *A simple Method for the Analysis of Binary Data.* Biometrics **48**: 577 – 585.

Shoukri, M. M., Martin, S. W. (1992). *Estimating the Number of Clusters for the Analysis of Correlated Binary Response Variable from Unbalanced Data.* Stat Med **11**: 571 – 760.

Snedcor, G. W., Cochran, W. G. (1967). *Statistical Methods (6th Edition).* Iowa, U.S.A.

Spilker, B. (1991). *Guide to clinical trials.* Raven Press. New York.

Sullivan, L. M., Dukes, K. A., Losina, E. (1999). *An introduction to hierarchical linear modeling.* Statist. Med. **18**: 855 - 888

Thompson, S. G., Pyke, S. D. M., Hardy, R. J. (1997). *The design and analysis of paired cluster randomised trials: An application of meta-analysis techniques*. Stat Med **16**: 2063 – 2079

Turner, R. M., Omar, R. Z., Thompson, S. G. (2001). *Bayesian methods of analysis of cluster randomized trials with binary outcome data.* Statistics in Medicine **20**: 453 - 472

Wears, R. L. (2000). *Statistical Methods for Analyzing Cluster and Cluster-Randomised Data*. Society for Academic Emergency Medicine; San Francisco, CA

Wikipedia, *The Free Encyclopedia*. *Paradox.* http://en.wikipedia.org/wiki/Paradox

Woolson, R. F., Bean, J. A., Rojas, P. B. (1986). *Sample Size for Case-Control Studies Using Cochran's Statistic.* Biometrics. **42**: 927 - 932

Zar, J. (1996). *Biostatistical Analysis (3rd edition).* Prentice-Hall International. Upper Saddle River

Zhang, J., Boos, D. D. (1996). *Mantel-Haenszel Test Statistics for Correlated Binary Data*. Biometrics 1997; **53**: 1185 – 1198